Caracterización de cultivos de maíz usando enfoque Clusterwise para la optimización de su rendimiento basado en la Mejor Búsqueda Armónica Global

Darwin Fabián Muñoz Pérez
José Luis Rivera Ibarra

Director: Est. MSc. (c) Hugo Andrés Dorado Betancourt
Codirector: PhD. Carlos Alberto Cobos Lozada

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Grupo de I+D en Tecnologías de la Información (GTI)
Línea de Investigación: Sistemas Inteligentes y Gestión de la Información
Popayán, diciembre de 2018

TABLA DE CONTENIDO

Lista de Tablas

Lista de Figuras

LISTA DE ANEXOS DIGITALES

Anexo 1: Artículo “Caracterización de cultivos de maíz usando el enfoque Clusterwise”.

Anexo 2: Artículo “Optimización del rendimiento en cultivos de maíz basado en la Mejor Búsqueda Armónica Global apoyada en Clusterwise”.

Anexo 3: Archivos iniciales.

Anexo 4:.Procesamiento datos climáticos en el CIAT.

Anexo 5:.Analisis archivos iniciales.

Anexo 6:.Procesamiento de archivos iniciales.

Anexo 7:.Seleccion de atributos con Wrapper.

Anexo 8:.Codigos fuente.

Anexo 9:.Resultados de la prueba no paramétrica de Wilcoxon.

Anexo 10: Resultados de la prueba t para muestras apareadas.

Anexo 11:.Analisis descriptivo del macro modelo.

Anexo 12:.Resultados del proceso de optimización.

Anexo 13: Ejecutables.

Anexo 14: Resultados pruebas adaptación enfoque clusterwise.

RESUMEN

La agricultura específica por sitio (AEPS) plantea la identificación de prácticas agronómicas mejor definidas, partiendo de condiciones espaciales y temporales presentadas en las zonas a ser usadas en la siembra. En el departamento de Córdoba – Colombia se genera en promedio los mayores niveles de producción de maíz, Córdoba tiene una superficie de alrededor de 23.980km^2 y se encuentra limitado con el mar Caribe y los departamentos de Sucre, Bolívar y Antioquia. Por su ubicación y extensión presenta una amplia variedad de condiciones climáticas y de tipos de suelos, por lo que orientar a los agricultores con prácticas genéricas, como se hace tradicionalmente, puede estar impidiendo alcanzar niveles óptimos de productividad.

Al conocer la variabilidad presentada al interior del departamento de Córdoba se presenta la necesidad de identificar zonas con características similares, para situaciones como la anterior se planteó el enfoque clusterwise, este es un método de agrupamiento que permite clasificar observaciones en grupos con características similares y para estos últimos, identificar modelos de regresión que expliquen el comportamiento particular de cada agrupación. Este enfoque a la fecha del presente trabajo de investigación no ha sido utilizado en el campo de la agricultura, por lo que el presente trabajo centra sus esfuerzos en implementar dos adaptaciones de este enfoque, utilizando las metaheurísticas de recocido simulado multi arranque (MSSA) y procedimiento de búsqueda codiciosa aleatorizada y adaptativa (GRASP), logrando identificar un número óptimo de zonas con características similares y sus respectivos modelos de regresión; permitiendo brindar orientación a los agricultores con prácticas de manejo más específicas.

Una vez identificadas las prácticas de manejo influyentes en las zonas con características similares se presenta la necesidad de conocer cómo desarrollar adecuadamente estas, con esto buscando aumentar los niveles de producción. Suponiendo que posiblemente aún no han sido identificadas estas prácticas por los agricultores del departamento de Córdoba, se adapta las metaheurísticas mejor búsqueda armónica global (GBHS) y optimización por enjambre de partículas (PSO), estas a partir de los grupos y modelos generados con la adaptación de clusterwise, logran identificar, que prácticas permiten a cada agricultor aumentar sus niveles de producción partiendo de la zona donde se realizara el cultivo.

CAPÍTULO 1

1 INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

El cultivo de maíz en Colombia es un sector muy importante dentro de la actividad agropecuaria [1], [2], en el que se ha observado por mucho tiempo una producción superior al millón de toneladas por año y un consumo que ha conseguido niveles próximos a los 20 kilos per-cápita [3], [4]. El maíz se considera una fuente de ingresos y de trabajo en varias regiones agrícolas de Colombia [5], [6]. Los cultivos se dan en diversos ambientes que están en alturas desde el nivel del mar hasta más de los 3.000 metros sobre el nivel del mar [4]. Además de ser una fuente de ingresos, es un alimento base para la población latinoamericana gracias a su importante contenido nutritivo y gran variedad de productos derivados [7].

Teniendo en cuenta la importancia de la producción de maíz a nivel nacional, los agricultores colombianos tienen el desafío de aumentar el rendimiento y la calidad de este producto, debido a que actualmente la producción nacional solo logra cubrir el 85% [8] del consumo humano en Colombia y se requiere de la importación del 15% restante, así como, para cubrir el resto de la demanda que incluye la alimentación de animales y el uso industrial. De la misma manera se proyecta que el crecimiento continuo de la población siga ocasionando un aumento en la demanda de este producto con el paso del tiempo [3], [9].

Se debe agregar que de acuerdo a estudios realizados por la Federación Nacional de Cultivadores de Cereales y Leguminosas (FENALCE) [8], las importaciones de maíz realizadas para cubrir la demanda faltante se hacen mediante tratados de libre comercio (TLC) con diferentes países como Estados unidos y México principalmente [3]. Sin embargo, esta solución ha afectado a los productores nacionales debido al costo competitivo del maíz importado.

Por otro lado, FENALCE junto con el apoyo de investigadores han logrado fortalecer conocimientos sobre el cultivo de maíz, permitiendo distinguir prácticas agrícolas más tecnificadas, aunque muchas veces de manera genérica, basados principalmente en conocimientos de expertos o prácticas tradicionales; pero, aún con poco uso de la Agricultura Específica por Sitio (AEPS). Para los agricultores de maíz esta situación actual podría implicar un pobre entendimiento de las técnicas a desarrollar durante la siembra, que a su vez podría llevar a un uso equivocado de los recursos, baja productividad, costos elevados de producción y dificultades para competir con el mercado internacional [10]. Hay que mencionar que, debido a la distancia geográfica entre algunos agricultores con cultivos en las mismas condiciones biofísicas, es difícil que se reúnan para socializar, compartir las técnicas con las que están obteniendo mejores resultados y de paso definir las mejores prácticas en condiciones similares a través de sus experiencias.

Ahora bien, definir los factores y técnicas más influyentes en el rendimiento de un cultivo de ciclo corto (cultivos de ciclo de vida inferior a los 365 días, con la particularidad que una vez se cosechan existe la necesidad de volverlos a sembrar) que está presente en diversas regiones es una labor compleja, ya que la producción se lleva a cabo en diferentes alturas, condiciones variadas de clima y suelo [4], [11]. Lo anterior implica que diversos factores pueden influir sobre la calidad y cantidad de la producción, siendo este el caso del departamento de Córdoba, que es considerado uno de los principales productores de maíz en Colombia [12], aportando cerca del 11.4% del total de la producción nacional con sus cultivos de maíz tradicional y tecnificado [5], [6] y del cual se conoce que no es un departamento con condiciones uniformes, debido a que se encuentra dividido en regiones, gracias a su topografía plana y montañosa que lo caracteriza como un departamento con clima sectorizado con variaciones entre los 18ºC y 28ºC.

De lo anterior, han surgido estudios centrando esfuerzos en el entendimiento de los elementos o decisiones más influyentes en los cultivos, indicando que cada vez que un agricultor realiza una siembra, hay un evento único [13], ya que es un experimento que prácticamente no tiene antecedentes, el cual se puede monitorear a lo largo de su desarrollo y producción. Con el registro de un número significativo de estos eventos y la implementación de un enfoque de modelación estadística tradicional (Generación de un modelo matemático basado en datos que se supone tiene forma de un modelo estocástico) o modelación basada en inteligencia artificial (Generación de un modelo matemático basado en datos sin hacer suposiciones sobre la distribución y comportamiento de los datos), puede proveerse información a los productores sobre cómo elegir técnicas para el manejo adecuado de sus cultivos [14]–[16].

Por otro parte, existen investigaciones donde han aplicado técnicas de conglomerados o agrupaciones (clustering), para la identificación de mejores prácticas de manejo agrícola, basándose en agrupaciones por propiedades de clima y suelo sin importar la ubicación geográfica donde se han sembrado diferentes cultivos como: maíz y plátano [17], [18]. Con estos análisis de clúster se ha logrado caracterizar el ambiente donde están estos cultivos, con el objetivo de identificar los factores de mayor contribución dentro de cada agrupación detectada.

Con los antecedentes expuestos, se puede observar que actualmente los agricultores de maíz tienen la posibilidad de aumentar la producción para suplir la demanda nacional de este producto en un porcentaje mayor. Una forma de aumentar los niveles de producción se consigue mediante la adaptación de prácticas de siembra basadas en la AEPS, logrando con esto tomar decisiones de prácticas de manejo acordes a las distintas condiciones específicas de cada sitio utilizado para el cultivo.

En la actualidad los investigadores a menudo precisan ajustar los modelos de regresión a un grupo de datos que puede ser no homogéneo con respecto a los factores involucrados [19], esto en un intento de identificar la relación entre una variable llamada dependiente y una o más variables llamadas independientes,

generalmente al momento de registrar eventos de agricultura desarrollados desde diversas zonas geográficas y asociados a varios agricultores, se obtienen datos no homogéneos, al dividir este conjunto de datos en sub conjuntos donde la información contenida sea homogénea se identifica las relaciones entre las variables y como estas varían dentro de cada sub conjunto.

Un posible aporte desde la ciencia de la computación para la agricultura específica por sitio, puede realizarse con la integración de lo trabajado hasta la fecha (análisis por conglomerado y modelos de regresión) usando el enfoque clusterwise, que logra agrupar los datos en clústeres con observaciones homogéneas y definir modelos de regresión dentro de cada clúster de forma simultánea. Este enfoque ha sido implementado en áreas como la gestión de pavimentos donde se han obtenido muy buenos resultados para la gestión vial [20]. En la AEPS para cultivos de ciclo corto se tiene un contexto similar al de la gestión de pavimentos, debido a que en ambos casos se usan muchas variables explicativas y sus contribuciones a la variable de respuesta varían de un sitio específico a otro.

Por lo anterior, en este trabajo se abordó la siguiente pregunta de investigación: ¿Cómo adaptar el enfoque clusterwise para caracterizar las zonas utilizadas en la siembra de maíz al interior del departamento de Córdoba – Colombia, a partir de datos históricos provistos por FENALCE y el CIAT[1], y con ello buscar optimizar el rendimiento (producción en kg por hectárea) en dichos cultivos?

1.2 APORTES DEL PROYECTO

La contribución de este proyecto desde la perspectiva de investigación se centró en generar nuevo conocimiento relacionado con la adaptación del enfoque clusterwise para la caracterización de las zonas utilizadas en la siembra del maíz que tienen condiciones similares de suelo, clima, entre otros. Para adaptar el enfoque clusterwise se realizó la evaluación y comparación de dos algoritmos metaheurísticos, Recocido Simulado Multi Arranque (Multi-Start Simulated Annealing, MSSA) y el procedimiento de búsqueda codiciosa aleatorizada y adaptativa (Greedy Randomized Adaptive Search Procedure, GRASP). Adicionalmente, se definieron dos algoritmos metaheurísticos basados en la mejor búsqueda armónica global (Global-Best Harmony Search, GBHS) y en Optimización por Enjambre de Partículas (Particle Swarm Optimization, PSO) para optimizar los modelos generados (seleccionando los atributos más importantes) sobre cada uno de los grupos, aspecto del que no se tiene referencia en el estado del arte. Con esto se buscó extraer las prácticas de manejo más influyentes en cada grupo identificado, con el fin de replicarlas en futuras siembras y optimizar los niveles de producción (rendimiento).

Desde la perspectiva de innovación, el proceso de generación de modelos por clusterwise y el proceso de optimización se ha empaquetado en una aplicación

[1] El Centro Internacional de Agricultura Tropical (CIAT) es una organización que realiza investigación colaborativa para mejorar la productividad agrícola y el manejo de los recursos naturales en países tropicales y en vía de desarrollo.

software desarrollada en el lenguaje de programación java la cual queda disponible bajo licencia GPL a la comunidad académica, científica e industrial que lo requiera. Se espera que su uso por parte de técnicos y tomadores de decisiones tenga un impacto positivo en los agricultores de maíz pertenecientes al departamento de Córdoba y que su uso sea replicado a nivel global en las actividades agrarias sin importar el cultivo.

Desde la perspectiva social, se espera que el uso de los modelos obtenidos beneficiara a los agricultores de maíz al interior del departamento de Córdoba, al permitirles conocer y replicar las mejores prácticas que inciden en los niveles de producción de maíz (kg/hectárea), gracias a la identificación y agrupación de zonas de cultivo con características similares tanto en clima como suelo, en las que se llevan a cabo distintas prácticas de manejo agronómico sin importar la distancia geográfica de dichas zonas.

1.3 OBJETIVOS

A continuación, se presentan los objetivos del proyecto conforme fueron aprobados por el Consejo de la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca en el documento del anteproyecto.

1.3.1 Objetivo general

Adaptar el enfoque clusterwise para soportar la caracterización de las zonas utilizadas en la siembra de maíz en el departamento de Córdoba – Colombia y buscar la optimización de rendimiento usando una metaheurística.

1.3.2 Objetivos específicos

- Definir una vista minable[2] sobre cultivos de maíz apoyándose en las fases de comprensión del negocio, compresión de datos y preparación de datos de CRISP-DM[3] sobre los datos subministrados por el CIAT y FENALCE, y el conocimiento de expertos de las mismas instituciones, buscando la caracterización de las zonas utilizadas para dichos cultivos al interior del departamento de Córdoba.
- Adaptar el enfoque clusterwise apoyándose en las fases de modelado y evaluación de la metodología CRISP-DM, para obtener la caracterización de las zonas utilizadas en los cultivos de maíz sobre la vista minable previamente definida y la comparación del desempeño de dos metaheurísticas (Recocido Simulado Multi Arranque y GRASP) que serán usadas para dar soporte en la distribución de los grupos (clústeres) de las zonas de cultivo con características similares.

[2] Tabla única desnormalizada compuesta por todas variables relevantes para un problema de minería de datos.

[3] CRISP-DM (Cross Industry Standard Process for Data Mining) es una metodología para desarrollar proyectos de minería de datos compuesta por 6 fases, en donde paso a paso describe las tareas a tareas para el desarrollo del proyecto además de brindar las pautas para el correcto manejo de los riesgos, los recursos entre otros.

- Adaptar la metaheurística de la Mejor Búsqueda Armónica Global (GBHS) buscando optimizar el rendimiento en los grupos previamente definidos, además de desplegar esta funcionalidad a través de una aplicación WEB, apoyándose en la fase de despliegue de la metodología CRISP-DM y haciendo uso del marco de trabajo SCRUM[4] para el desarrollo de la misma.

1.4 RESULTADOS OBTENIDOS

A continuación, se presenta un resumen de los resultados más destacados que se obtuvieron en este trabajo de grado:

- **Monografía de trabajo de grado:** Se refiere al presente documento donde se expone el estado del arte del problema y se realiza una descripción detallada del proceso de investigación seguido y de los algoritmos adaptados, siguiendo con los resultados de la experimentación y la comparación de los algoritmos, finalizando con las conclusiones del trabajo realizado y lo que el grupo de investigación espera realizar en el corto plazo en relación con la temática abordada.
- **Aplicación software:** Aplicación web que presenta un formulario en el que el usuario puede escoger datos de demostración e interactuar con estos para solicitar su optimización, esta aplicación se entrega con video demostrativo del uso de la aplicación web.
  Aplicación de escritorio para la caracterización de cultivos usando el enfoque clusterwise, esta aplicación se entrega con su respectivo manual y video demostrativo del uso.
  Ambas aplicaciones se entregan junto con su respectivo código fuente.
- **Artículo 1**: un artículo resumen del procesamiento de los datos hasta la construcción de la vista minable y el proceso de adaptación del enfoque de clusterwise al cultivo de maíz en proceso de evaluación con la siguiente referencia preliminar: Carlos Cobos, Darwin Fabián Muñoz, Hugo Andrés Dorado, José Luis Rivera, “Caracterización de cultivos de maíz usando el enfoque Clusterwise”, que se encuentra en proceso de evaluación en la revista Computers and Electronics in Agriculture. Ver **ANEXO 1**.
- **Artículo 2**: un artículo resumen del procesamiento de optimización de los cultivos incorporando el concepto de Agricultura Especifica Por Sitio y su despliegue a través de una aplicación Web con la siguiente referencia preliminar: Carlos Cobos, Darwin Fabián Muñoz, Hugo Andrés Dorado, José Luis Rivera, “Optimización del rendimiento en cultivos de maíz basado en la Mejor Búsqueda Armónica Global apoyada en Clusterwise”, Ver **ANEXO 2**.

[4] SCRUM es un marco de trabajo ágil con un enfoque iterativo e incremental para gestionar el desarrollo de productos en el cual se pueden emplear diferentes procesos y técnicas. Es un proceso de equipo con unos roles definidos, donde el desarrollo se divide en iteraciones (Spring) y a lo largo de cada iteración se realizan una serie de eventos, que ayudan a la mitigación de riegos y la rápida adaptación a cambios.

1.5 ORGANIZACIÓN DEL DOCUMENTO

A continuación, se introducen los temas que se tratan en cada uno de los siguientes capítulos de la presente monografía.

Capítulo 2: Contexto teórico y estado del arte. En este capítulo se presentan los conceptos necesarios para la comprensión del enfoque clusterwise, los algoritmos utilizados en este trabajo y que son usados comúnmente para la solución de problemas combinatoriales y trabajos previos relacionados con sistemas de apoyo a la toma de decisiones en agricultura.

Capítulo 3: Definición de la vista minable. Se describe el tratamiento realizado a los registros sobre cultivos al interior del departamento de Córdoba. Se muestra como el proceso seguido permitió obtener un único archivo o tabla con la información pertinente para realizar posteriormente el proceso de modelado (adaptación del enfoque de clusterwise) y optimización.

Capítulo 4: Adaptación del enfoque clusterwise a cultivos de maíz. Este capítulo explica la adaptación del enfoque clusterwise mediante el acoplamiento de algoritmos metaheurísticos como MSSA y GRASP. Describe además las pruebas estadísticas sobre los resultados obtenidos y finalmente las observaciones proporcionadas por los expertos del área de investigación en Decisión de Políticas y Análisis (DAPA) perteneciente al CIAT, todo lo anterior con el fin de poder usar el enfoque en cultivos de maíz al interior del departamento de Córdoba.

Capítulo 5: Optimización. Se presenta el ajuste realizado a las metaheurísticas GBHS y PSO que se soportan en los modelos generados por la propuesta de adaptación de clusterwise para la optimización del rendimiento en los cultivos; esto es, definir las mejores técnicas según AEPS para obtener más rendimiento (cantidad de maíz en toneladas por hectárea) en los cultivos, Describe además el proceso ejecutado para desarrollar la aplicación web que permite realizar la optimización en una interfaz gráfica de usuario.

Capítulo 6: Conclusiones y trabajos futuros. En este capítulo se presentan las conclusiones obtenidas a partir de los resultados de los experimentos y del desarrollo como tal del proyecto. Además, los trabajos futuros que el grupo de investigación espera desarrollar en la temática del proyecto.

Por último, se presentan enumeradas todas las referencias bibliográficas utilizadas en el desarrollo del presente trabajo de grado.

CAPÍTULO 2

2 CONTEXTO TEÓRICO Y ESTADO DEL ARTE

2.1 CONTEXTO TEÓRICO

La presente sección del documento centra su atención en mostrar los conceptos más significativos relacionados con el presente proyecto de investigación, esto con la intención de proveer claridad en la temática que se aborda. Para mayores detalles, se recomienda recurrir a las referencias bibliográficas presentadas a través del capítulo.

2.1.1 Clustering

Esta tarea de minería de datos tiene como objetivo agrupar observaciones de acuerdo con un criterio, por lo general este criterio es una medida de distancia o similitud entre las observaciones (registros o tuplas), tal medida puede definirse de diferentes maneras como, por ejemplo, la distancia euclidiana, la distancia manhattan, la distancia Gower, la similitud de cosenos, entre otras [21].

2.1.2 Clusterwise

Este enfoque propone trabajar la agrupación de manera entrelazada con la creación de modelos de regresión de cada uno de los grupos que se evalúan. Es decir, partiendo de un conjunto de datos con *M* observaciones, pretende encontrar un óptimo número de clústeres, donde cada uno de estos tendrá su propio modelo de regresión y la calidad de los modelos de regresión obtenidos define la calidad de solución de agrupamiento.

Lo anterior se logra mediante una propuesta de intercambio que arranca con una repartición inicial de las *M* observaciones entre los *N* clústeres y en cada iteración se realiza intercambio de observaciones entre los *N* clústeres, luego se generan los respectivos modelos de regresión, buscando por ejemplo que la suma global de los errores al cuadrado de los modelos de regresión sobre los clústeres se minimice [22], [23].

2.1.3 Estadística

La estadística es un método científico que, a través de la recolección, organización, y análisis de datos, busca generar conclusiones que faciliten la toma de decisiones. Para realizar el análisis de los datos normalmente se deben evaluar algunos supuestos sobre ciertas propiedades de los mismos como linealidad, normalidad e independencia, siendo esto una limitación importante, debido a que en ocasiones los datos son diversos y no son homogéneos, por lo tanto, no se pueden describir mediante distribuciones de probabilidad conocidas [24], [25].

2.1.4 Inteligencia artificial

La inteligencia artificial intenta simular la esencia y la naturaleza de la vida mediante la construcción de sistemas artificiales que resaltan ciertas propiedades del comportamiento de organismos vivos, bien sea de manera individual o colectiva. Lo anterior con el fin de establecer métodos de búsqueda de soluciones óptimas dentro de un gran número de posibles soluciones al problema en cuestión [26].

2.1.5 Algoritmos metaheurísticos

Los algoritmos metaheurísticos son métodos de inteligencia artificial que permiten encontrar soluciones óptimas o cercanas al óptimo en problemas complejos (continuos, discretos o binarios), con un costo computacional razonable y en un periodo de tiempo aceptable en comparación con los métodos exactos [27]. Los problemas complejos se caracterizan por contar con un espacio de búsqueda grande, que crece exponencialmente en relación con los parámetros del problema, y por ello se busca encontrar la mejor solución posible [28], sin garantía de encontrar la solución óptima global al problema [29]. Los métodos exactos o que realizan búsquedas exhaustivas pueden garantizar la solución óptima pero a un costo computacional que en muchos casos los hacen inviables [27].

Una de las principales ventajas de las metaheurísticas frente a otros métodos se centra en su gran flexibilidad, ya que estas pueden ser usadas para resolver una gran variedad de problemas [27]. Las metaheurísticas basan su funcionamiento en dos conceptos principales: diversificación (exploración) e intensificación (explotación). El proceso de diversificación se encarga de explorar el espacio de búsqueda global y seleccionar la(s) mejor(es) solución(es) encontrada(s). El proceso de intensificación se especializa en la búsqueda de la(s) solución(es) candidata(s) en una región local, partiendo del hecho que se ha encontrado una buena solución en dicha región [29].

Los algoritmos metaheurísticos se pueden clasificar de diversas maneras, entre ellas: basados en poblaciones, los cuales utilizan un conjunto de individuos que evolucionan o se mejoran durante su ejecución; y los de estado simple, que usan un único individuo y lo evolucionan buscando explotar zonas prometedoras y saltar de una región a otra del espacio de búsqueda para hacer exploración y salir de óptimos locales.

2.1.5.1 Recocido Simulado

El recocido simulado (Simulated Annealing, SA) es un algoritmo que se inspira en el proceso de enfriamiento aplicado a los metales y las cerámicas [27]. Para esto se utiliza la representación de un único individuo, el cual inicialmente explora el espacio de búsqueda y con el paso de las iteraciones se dedica a explotar una región del espacio de búsqueda, pero haciendo pequeños saltos de exploración. Lo anterior, se logra gracias a que se tiene un parámetro de temperatura y una función que dependiendo del valor de la temperatura y de la calidad de las nuevas soluciones generadas, las acepta con una cierta probabilidad si no son consideradas como buenas, permitiéndole así salir de óptimos locales. Cuando la temperatura ya tiene

un valor muy bajo la probabilidad de aceptar este tipo de soluciones de mala calidad es mínima, es ahí donde se centra sólo en explotar el espacio de búsqueda que es cercano a la solución que en ese momento está considerada como óptima [27].

Con la finalidad de poder mejorar la solución reportada como óptima, al final de la ejecución del recocido simulado, se adopta la acción de elevar el valor de la temperatura nuevamente y reiniciar el algoritmo, tomando el óptimo como el estado inicial del individuo para la nueva ejecución, esta operación se realiza en repetidas ocasiones hasta satisfacer un criterio de parada, este procedimiento es conocido como multi arranque [30].

2.1.5.2 GRASP

El procedimiento de búsqueda codiciosa aleatorizada y adaptativa (Greedy Randomized Adaptive Search Procedure, GRASP), es una metaheurística de estado simple, que realiza la búsqueda siguiendo una trayectoria a partir de diversas soluciones iniciales [31].

Cuenta con una etapa inicial de preprocesamiento en la que se define la representación del individuo y luego se realiza un proceso iterativo que termina cuando se cumple con un criterio de parada, este puede ser: un cierto nivel de calidad en la solución encontrada, un número de iteraciones o un máximo tiempo de ejecución.

Al interior del proceso iterativo se cuenta con dos fases: construcción y optimización local, la de construcción se apoya en un procedimiento greedy aleatorizado y soportado en una lista restringida de candidatos (Restricted Candidate List) para construir una solución de buena calidad; y luego con el procedimiento de optimización local (explotación) se refina esta solución buscando en su vecindario soluciones que tengan mayor calidad. Al final de cada iteración se compara la solución construida por estas etapas con la mejor encontrada hasta el momento, estableciendo cuál presenta una mejor medida de calidad [31].

2.1.5.3 PSO

La optimización por enjambre de partículas (Particle Swarm Optimization, PSO) es un algoritmo de búsqueda poblacional, bio inspirado en el comportamiento social observado en las parvadas, manadas y cardúmenes, donde se toman decisiones a partir del conocimiento colectivo [32]. Para representar esta conducta la metaheurística trabaja con un conjunto de individuos (enjambre) los cuales representan múltiples soluciones (partículas) al problema en cuestión.

El algoritmo inicia definiendo el enjambre de partículas, posteriormente estas empiezan a desplazarse por el espacio de búsqueda actualizando su posición y evaluando la calidad de su trayectoria, buscando ubicarse en lugares más prometedores, este proceso se repite hasta cumplir un criterio de parada.

El proceso de desplazamiento sobre el espacio de búsqueda se realiza con el objetivo de encontrar una solución óptima en un tiempo razonable, esta búsqueda

se apoya en un componente que define su desplazamiento y su dirección (vector de velocidad). Este módulo está compuesto por el conocimiento de la partícula (mejor posición encontrada) y el conocimiento colectivo (mejor posición encontrada por el enjambre) [32].

2.1.5.4 GBHS

La mejor búsqueda armónica global (Global-Best Harmony Search, GBHS), es una metaheurística poblacional, inspirada en el proceso de improvisación musical [34]. El algoritmo se divide en las etapas de inicialización, improvisación y actualización de la memoria armónica, con las cuales se busca encontrar una solución óptima dentro del espacio de búsqueda. La memoria armónica es una lista que representa un conjunto de soluciones donde cada una es una interpretación única del problema que se busca optimizar.

En la etapa de inicialización se construye un conjunto de *N* soluciones (armonías) y se almacenan en la memoria armónica (población de un algoritmo genético o enjambre de PSO), con esto se inicia un proceso iterativo en donde se realiza la improvisación de partituras (construcción de una nueva solución), luego se entra a evaluar y en caso de identificar que la improvisación presenta una mejor medida de calidad que la peor solución de la memoria armónica, se procede a realizar el remplazo conservando la nueva melodía, en caso contrario la memoria armónica no se altera. El algoritmo continúa improvisando distintas soluciones hasta cumplir con un criterio de parada [33], [34].

2.1.6 K-means

K-means es un algoritmo para clustering (aprendizaje no supervisado) que busca agrupar un conjunto de datos a partir de su estructura interna (propiedades, características) [35]. La asignación de un registro a un determinado clúster/grupo se realiza buscando establecer la menor distancia (mayor similitud) entre los registros y el centroide de cada clúster.

El algoritmo presenta tres etapas; en la primera se realiza una asignación inicial de *N* centroides a cada uno de los *N* clústeres, normalmente en forma aleatoria, con esto se da inicio a un proceso iterativo dentro del cual se realiza dos etapas principales, la asignación de membrecías y la actualización de centroides. La asignación de membrecías se encarga de repartir los datos entre los *N* grupos, asignando cada dato al grupo donde su centroide (representante del grupo) tenga la menor distancia. Luego de repartir los registros entre los grupos, se pasa a la fase de actualización de centroides, en la cual se promedian todas las dimensiones de los datos de cada grupo para recalcular el centroide que los representa. Con esto se concluye una iteración y se procede a la siguiente repitiendo los pasos 2 (asignación de membrecías) y 3 (actualización de centroides) hasta cumplir con un criterio de parada [35].

2.2 ESTADO DEL ARTE

La revisión de la literatura señala que para optimizar los cultivos se han propuesto diferentes métodos basados en la estadística y en la inteligencia artificial, en algunos casos incluyendo el clustering. Aunque el clusterwise no se ha usado en la optimización de cultivos, en esta sección se incluye un trabajo que se tomó como base para la presente investigación.

2.2.1 Enfoques de modelación estadística tradicional y basada en inteligencia artificial

Estos dos enfoques (estadística clásica e inteligencia artificial) son los que se han usado recientemente en el estado del arte. A continuación, se presentan los trabajos más relevantes al presente proyecto organizados cronológicamente.

En el 2006 [36], partiendo de que no se habían realizado estudios para comprender la variabilidad de la calidad del maíz con enfoques estadísticos no lineales, se desarrolló un proyecto con el propósito de ilustrar el uso de redes neuronales artificiales (RNAs) para identificar los factores más influyentes en el rendimiento del maíz y la calidad del grano (contenido de proteína y el peso del grano). En este estudio se utilizaron datos de siembras realizadas en dos campos al este de Illinois (Estados Unidos) en el año 2000, los cuales estaban compuestos por diferentes características, destacándose: el contenido proteínico del grano (muestra aleatoria de semillas), el rendimiento del maíz (Kg/ha) tomado a partir de un monitor de rendimiento y características del suelo como materia orgánica, Bray-P, potasio, capacidad de intercambio catiónico, pH, entre otras. El estudio se realizó por separado para los dos campos, haciendo uso de estadística descriptiva y multivariada con el fin de entender mejor las relaciones entre las diferentes variables y reducir la redundancia del conjunto de datos, además de seleccionar las variables más importantes para su posterior análisis. Las variables seleccionadas se examinaron mediante RNAs, y usando Intelligent Problem Solver (IPS) se logró encontrar las mejores combinaciones de variables, algoritmos de entrenamiento y parámetros asociados para obtener resultados óptimos. Adicionalmente se dividió el conjunto de datos en tres subconjuntos al azar definiendo los subconjuntos de formación (entrenamiento), selección (validación) y pruebas. Se definieron varios modelos y por último se seleccionó el mejor, al cual se le realizó un análisis de sensibilidad para evaluar la importancia relativa de cada variable en la explicación del rendimiento del maíz y la calidad del grano, obteniendo al final una selección de 7 y 8 factores relevantes de los campos 1 y 2 que explican la variabilidad del rendimiento y la calidad con un porcentaje entre el 68% al 92% y el 68% al 99% respectivamente para los dos campos.

En 2008 [37], se realizó un estudio sobre cultivos de maíz al oeste de Kenia, con el objetivo de investigar la importancia de la fertilidad del suelo y factores de manejo respecto a la variabilidad del rendimiento de pequeños cultivos familiares y comerciales. Estos cultivos estaban ubicados en las regiones con mayor potencial agrícola como Aludeka, Emuhaya y Shinyalu. En este estudio se recolectó información sobre rendimiento y manejo agronómico de cultivos en 159 campos

pertenecientes a 60 granjas. Debido a la complejidad del conjunto de datos se optó por utilizar árboles de clasificación y regresión (CART) y se agruparon las variables predictoras en tres categorías: general, manejo y suelo-paisaje. De estas categorías se destacan variables relacionadas con el estudio del suelo, además de información relacionada con el cultivo como su ubicación, fertilidad (clasificación de fertilidad según criterio del agricultor), intensidad de uso de recursos (tipos de abono y cantidades, frecuencia de limpieza, densidad de siembra, entre otros), retraso de siembra (retraso frente a las fechas recomendadas en Kenia para siembra) y nivel de maleza. El algoritmo CART se utilizó junto con validación cruzada de 10 carpetas para obtener modelos predictivos más robustos, usando como variable objetivo el rendimiento del maíz (grano, biomasa, rendimiento de grano por planta y biomasa por planta), este algoritmo también se ejecutó dos veces con el fin de obtener dos modelos con base en diferentes conjuntos de variables predictoras. Para el primer modelo se escogieron las categorías general y manejo, y para el segundo modelo se incluyeron todas las categorías, obteniendo relaciones similares en ambas ejecuciones. Además, se encontraron diferentes relaciones asociadas a la influencia de uso de recursos y fechas de siembra. Los resultados destacaron que en los campos categorizados según su fertilidad como fértiles, más del 50% de los agricultores invierten más recursos, mientras que en los campos catalogados como no fértiles los cultivos se siembran con desfase sobre las fechas de siembra recomendadas y no se invierten recursos influyendo notablemente sobre el rendimiento, además la variación del rendimiento en estos campos asciende al 100% respecto al rendimiento ideal para estas zonas. Finalmente, el estudio determina que los campos catalogados como no fértiles deben ser objeto de importantes estrategias de rehabilitación para mejorar la productividad de la tierra y los medios de vida rurales al oeste de Kenia ya que estos campos ocupan la mayor parte del área agrícola.

En el 2009 [16], se analizaron 488 registros de producción de mora Andina en Colombia, estos datos se componían de información sobre clima, suelo, manejo y rendimiento de estos cultivos. La motivación principal de este estudio surgió debido a que el sector estaba caracterizado por agricultores con limitada información sobre los factores que influyen en la producción debido a que este cultivo es poco investigado. Dentro del análisis se realizó una identificación de las variables más relevantes sobre la productividad final, para lo cual se emplearon metodologías basadas en RNAs (perceptrón multicapa) y análisis de sensibilidad. Como resultado se obtuvo que para este cultivo: la profundidad del suelo, la temperatura media, el drenaje externo y la precipitación acumulada del primer mes antes de la cosecha, presentaron mayor influencia sobre la productividad. Después se identificó un total de 6 grupos utilizando análisis clustering dentro de los cuales se distribuyeron los registros, con esto se encontraron zonas con comportamientos similares en niveles de producción que fueron utilizadas en los mapas auto organizados (red neuronal de Kohonen) para visualizar las relaciones entre la productividad y cada una de las variables identificadas, lo cual permitió hallar las condiciones óptimas que conducen a altas producciones, demostrando el gran aporte que tuvieron estas técnicas de análisis basadas en inteligencia artificial para un sistema tan limitado en datos.

En el 2011 se recolectaron 256 registros de siembras de lulo realizadas por agricultores en el departamento de Nariño – Colombia durante un periodo de dos años [14], con el objetivo de caracterizar el sistema de producción usando variables de gestión y condiciones ambientales de estos cultivos. Se procedió inicialmente a identificar los factores que contribuían más en la explicación del rendimiento, para lo cual se emplearon dos metodologías, una estadística y una computacional con el fin de comparar los resultados, se utilizó el modelo de regresión lineal con el método de stepwise y el modelo de perceptrón multicapa con una métrica de sensibilidad. Como resultado se obtuvo que ambas coincidieron con la elección de variables relevantes, pero el perceptrón identificó una variable más que la regresión lineal. Posteriormente, se implementó un mapa auto organizado (red neuronal de Kohonen) para hallar conglomerados con condiciones homogéneas, estos grupos se formaron a partir de las variables encontradas como más relevantes y con esto se identificaron tres grupos; finalmente se definió un modelo mixto que trabajó con los 3 grupos y variables de localización y sitio de la granja. Los resultados finales evidenciaron como factores de importancia en los cultivos de lulo: (I) ubicación y el efecto del medio ambiente, (II) asociación entre la producción y la posición geográfica que podrían relacionarse con prácticas o aspectos sociales y (III) las habilidades de gestión de explotaciones que utilizan la finca. Se concluyó que la implementación de diversas metodologías fue de gran ayuda para interpretar la información proporcionada por los agricultores para este cultivo.

En el 2012 [17], se realizó un acercamiento a la caracterización de las zonas cultivadas con plátano; esto, partiendo de la premisa de que existen ventajas sobre obtener información que es proporcionada directamente por los agricultores junto con datos sobre características climáticas y del suelo, siendo posible describir el sistema productivo bajo los fundamentos de la agricultura específica por sitio, apoyándose en métodos estadísticos multivariados. El proceso inició realizando un análisis de componentes principales (principal component analysis) por separado a las variables climáticas y a las características de suelo, esto debido a la naturaleza de los valores utilizados para representar la información. Para este proceso se utilizaron técnicas de análisis para variables numéricas y para variables categóricas, encontrando en un número reducido de variables un alto porcentaje explicativo frente a la variable de rendimiento del cultivo. El siguiente paso fue la identificación de conglomerados con condiciones homogéneas en las variables climáticas y por separado las variables del suelo, encontrando 10 y 5 grupos respectivamente, una vez identificados los grupos se caracterizaron a partir de los aspectos descriptivos de cada variable independiente, obteniendo las medidas de tendencia central, posición y variabilidad para las variables numéricas y las frecuencias y modas para las variables cualitativas dentro de cada grupo. Finalmente, se aplicó un nuevo análisis descriptivo sobre las variables de manejo y rendimiento, pero esta vez con la integración de los grupos conformados por clima y suelo, dando origen a un ambiente de producción específico que bajo estas características es llamado zona agroecológica.

En 2014 [38], se realizó un estudio sobre cultivos de maíz con el objetivo de determinar las características más influyentes en el rendimiento del grano, haciendo

uso de modelos de selección, agrupación y árboles de decisión. El estudio se desarrolló en un ambiente controlado ubicado en la Granja Experimental de la Facultad de Agricultura de la Universidad de Shiraz (Irán) en donde se recolectó información de dos experimentos de campo realizados entre los años 2008 y 2009, además de datos de literatura sobre la fisiología del maíz formando un conjunto de 166 registros con 22 variables (características fisiológicas y agronómicas). En este estudio se realizó selección de características, detención de anomalías, agrupación de registros (k-means) y clasificación con árboles de decisión (CART) logrando obtener diferentes resultados, entre los cuales se destaca que la selección de características tuvo un efecto positivo en la agrupación, pero fue irrelevante en los arboles de decisión. Con la selección de características se obtuvo un subconjunto de 12 características relacionadas con el rendimiento del grano de maíz como fecha de siembra-ubicación, peso seco del tallo, tipo de suelo, peso final del grano, contenido máximo de agua del grano, peso seco de la mazorca, entre otras. Finalmente, se logró determinar que CART tuvo un mejor desempeño y se construyó un árbol, considerando la característica siembra-ubicación (fechas de siembra recomendadas por país) como la más importante y usándola para crear los principales subgrupos y ramas del árbol. Además, las relaciones identificadas estuvieron alineadas con estudios previos como la relación estrecha entre el número de granos por mazorca que se alcanzan en la madurez y el peso del grano. Por otro lado, el estudio destaca que las técnicas usadas son herramientas útiles para que los fisiólogos de cultivos logren seleccionar los rasgos más importantes para el sitio y el campo individual de acuerdo con patrones fisiológicos y agronómicos de un cultivo.

En 2016 [39], se realizó un estudio sobre cultivos de arroz irrigado y arroz de secano ubicados en dos regiones de los departamentos del Tolima y Meta en Colombia, con el objetivo de analizar el papel de la variabilidad climática como un factor limitante para los cultivos de estas zonas. Además, para poder evaluar los diversos patrones climáticos sobre los que se produjo el arroz y así poder cuantificar el impacto en el rendimiento, e identificar las condiciones más adecuadas para estos cultivos. Este estudio introdujo un enfoque basado en datos observacionales y técnicas de minería de datos para apoyar la toma de decisiones en cultivos de arroz. Para esto, se realizó la recolección de 1615 registros de cultivos de arroz de los cuales 1240 fueron de arroz irrigado tomados desde el año 2007 a 2013 correspondientes a la zona del Tolima y 375 de arroz de secano tomados desde el año 2007 a 2014 correspondientes a la zona del Meta, los cuales se recolectaron de diferentes fuentes y adicionalmente, estos registros contenían información obtenida de estaciones meteorológicas. En este estudio se realizó una evaluación de diferentes algoritmos para clustering y se optó por usar un algoritmo basado en árboles, siendo Conditional Inference Forests (CIF) el más adecuado a los objetivos del estudio; además, se utilizó Kruskal-Wallis para realizar un postratamiento sobre las distribuciones del rendimiento de los conglomerados, para poder agrupar y distinguirlos entre sí. En CIF se utilizó ntree=2000 y una muestra aleatoria de 27 variables de entrada, formando un conjunto de 100 modelos y limitando el número de observaciones analizadas en un modelo a 500, debido a la carga computacional de este algoritmo. Para conservar la precisión y mejorar el proceso de agrupación

se utilizó distorsión dinámica de tiempo y finalmente se logró extraer una explicación de los datos donde los predictores de cada modelo lograron definir que la temperatura afectaba de una forma distinta a cada grupo en diferentes etapas de desarrollo para las siembras de arroz irrigado, así como también que la disponibilidad de agua en los cultivos de secano es primordial en la etapa vegetativa. De acuerdo con los modelos se logró obtener que los cambios climáticos influyen sobre el rendimiento en un 24.7% y un 33.1% en las zonas del Tolima y Meta respectivamente.

2.2.2 Clusterwise

Aunque no se encuentra una referencia de clusterwise directamente aplicada a cultivos de maíz, se precisa mencionar un trabajo previo que ha sido muy relevante para la presente investigación y está relacionado con la gestión de pavimentos. Dicho trabajo es del año 2017 y en este se analizaron 4138 muestras de pavimento distribuidas en 17643 registros que fueron recolectadas entre 2001 y 2012 en el estado de Nevada - Estados Unidos [20]. En busca de aportar en la identificación temprana de segmentos de pavimento que requieren mantenimiento y los tiempos apropiados para ejecutar dichas actividades y con esto minimizar costos de intervención debido a fallos sustanciales que requiere una rehabilitación o reconstrucción severa, se propuso un modelo de regresión por clusterwise. Este trabajo es actualmente el estado del arte en la gestión de pavimentos y corresponde a una tesis de doctorado, en el que se implementó un algoritmo basado en recocido simulado y una formulación matemática, para la identificación de un número correcto de grupos con características similares, donde cada grupo está conformado por *N* registros y cada registro está asignado a uno de los grupos identificados. La función objetivo usada por el algoritmo es el criterio de información bayesiano, este es utilizado para evaluar la exploración de las posibles combinaciones de variables independientes dentro de cada agrupación de segmentos y de esta forma seleccionar el modelo más apropiado para cada grupo. La exactitud predictiva de los modelos resultantes se evaluó utilizando errores cuadráticos medios, errores cuadráticos medios normalizados y la media de los errores absolutos. Los resultados de este trabajo evidenciaron que los modelos no lineales fueron más precisos que los modelos lineales en la estimación de la variable dependiente.

Tabla 1 . Tabla de resumen trabajos estado del arte

Referencia	Tema	Conjunto de datos	Técnica	Objetivo
[36]	Maíz	Características de suelo.	RNAs	Determinar influencia de diferentes factores en el rendimiento.
[37]	Maíz	Características de suelo y prácticas de fertilización.	CART	Determinar influencia de diferentes factores en el rendimiento.
[16]	Mora Andina	Características de clima, suelo y manejo.	RNAs y Clustering	Determinar influencia de diferentes factores en el rendimiento.
[14]	Lulo	Características de clima y manejo.	RNAs, Regresión lineal y Mapas de kohonen	Caracterización del cultivo e influencia de diferentes factores en el rendimiento.

Referencia	Tema	Conjunto de datos	Técnica	Objetivo
[17]	Plátano	Características de clima y suelo.	Clustering	Caracterización de zonas de cultivo.
[38]	Maíz	Características de suelo y manejo.	CART y Clustering	Determinar influencia de diferentes factores en el rendimiento.
[39]	Arroz irrigado y de secano	Características de clima	Clustering y Conditional inference forest	Determinar influencia de factores climáticos en el rendimiento.
[20]	Pavimentos	Características de clima, tráfico vehicular y condiciones del pavimento.	Clusterwise	Caracterizar tramos viales y determinar factores relevantes en la resistencia y durabilidad del pavimento.

CAPÍTULO 3

3 DEFINICIÓN DE LA VISTA MINABLE

Una vista minable es una única tabla donde se encuentra de manera desnormalizada la información que por lo general proviene de diferentes tablas de una o varias bases de datos o de diversos archivos, y se compone por todas las variables relevantes para un problema de minería de datos, además la información almacenada en esta ha sido previamente procesada para garantizar la veracidad y consistencia de los datos contenidos [40].

A continuación, se presenta el proceso que fue desarrollado para construir la vista minable, este se soportó en las etapas de compresión del negocio, compresión de datos y preparación de datos de la metodología CRISP-DM, y tomó como fuente de datos inicial un conjunto de archivos entregado por el CIAT y FENALCE (**ANEXO 3**).

3.1 COMPRENSIÓN DEL NEGOCIO

Los objetivos del negocio corresponden a lo definido en la sección 1.2 del presente documento.

La evaluación de la situación se desarrolló durante la escritura del anteproyecto, y se puede consultar en las secciones actualizadas 1.1 y 2.2 relacionados con el Planteamiento del problema y el Estado del Arte, esto se realizó con el objetivo de conocer la terminología del negocio, supuestos, requerimientos, como posibles riesgos del proyecto se identificaron la limitante de tiempo para su ejecución, así como también la disponibilidad de recursos (personal experto para validación y retroalimentación durante las diferentes fases).

El objetivo del proyecto de minería de datos corresponde al definido en sección 1.3.2, en donde se busca caracterizar las zonas de los cultivos de maíz al interior del departamento de Córdoba (Colombia).

El plan de proyecto se desarrolló en la escritura del anteproyecto y está definido en el cronograma de actividades.

3.2 COMPRENSIÓN DE DATOS

En noviembre de 2017 se realizó una estadía de investigación en las instalaciones del CIAT (sede Palmira, Valle del Cauca) por parte de los autores del presente trabajo (Darwin Fabián Muñoz Pérez y José Luis Rivera Ibarra) con el objetivo de conocer el proceso de recolección de datos realizado por el CIAT y FENALCE.

El conjunto de datos fue tomado a partir de seguimientos realizados a cultivos de maíz al interior del departamento de Córdoba – Colombia durante un periodo de

tiempo comprendido entre los años 2013 y 2017, los registros recolectados contienen información correspondiente a prácticas de manejo agrícola, características de suelo e información climática.

La recolección de datos se realizó mediante dos vías: la primera fue a través de la plataforma virtual **http://siria.fenalce.org/**, la segunda por medio de técnicos agrónomos de FENALCE, los cuales se desplazaron hasta las zonas de cultivo para realizar un levantamiento manual de información y posteriormente ingresar estos datos en la plataforma virtual.

Dentro de la información capturada se contó con 813 registros sobre caracterización de lotes; estos datos fueron capturados por los agricultores siguiendo la guía RASTA[5], en la mayoría de los casos acompañados por personal de FENALCE, en estos registros se permite conocer el área sembrada, nivel de inclinación, ubicación y una caracterización del terreno a ser utilizado en la siembra, esto gracias al procedimiento realizado con la mencionada guía, adicionalmente de los lotes se contó con información asociada a la localización como lo es las coordenadas geográficas, el municipio donde se encuentra ubicado y el nombre del lote.

La guía RASTA se desarrolló como una metodología para caracterizar el suelo y el terreno del lote a ser utilizado en las siembras de una manera simple, con el objetivo de que el agricultor pueda conocer mejor los recursos con los que cuenta y la forma correcta de utilizarlos, la guía analiza diferentes características del suelo como su forma, el color, la textura, la estructura, potencial de Hidrogeniones (pH), pedregosidad, capas endurecidas, moteados, resistencia al rompimiento y presencia de carbonatos. Luego de tener claro estas características la guía proporciona en una segunda parte una serie de observaciones de campo que debe tener en cuenta que se deberán relacionar con las características medidas inicialmente; finalmente eso le permitirá identificar cuatro propiedades del suelo, como son: la materia orgánica, el drenaje, la profundidad efectiva y la presencia de sales.

En cuanto a los datos del estado del clima durante el tiempo que duran los cultivos, estos se generaron a partir de las coordenadas de geo localización asociadas a los lotes, las fechas de siembra y datos de estaciones meteorológicas de la localidad obtenidas del IDEAM[6], esto mediante el proceso definido por el CIAT (**ANEXO 4**). Con los resultados de este proceso se permite conocer los promedios de la temperatura, niveles de radiación y porcentajes de precipitación, durante las etapas presentes en cultivos de maíz (vegetativa, formación, madurez).

Por último, se recolectaron 998 eventos de cultivo (esto quiere decir que para cada lote se registró 1 o más eventos de cultivo e inclusive se encontraron lotes que no

[5] Guía desarrollada por el CIAT, Universidad Nacional de Colombia y Corporación BIOTEC, que busca permitir a los agricultores de manera simple y sin mayores recursos, poder realizar la caracterización de suelos y terrenos.

[6] El Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) es una entidad del gobierno de Colombia dependiente del Ministerio de Ambiente y Desarrollo Sostenible.

contaron con dichos eventos) que permitieron apreciar las prácticas de manejo agrícola (acciones realizadas y su intensidad o frecuencia), con el registro de riego, fertilización, control de malezas y de siembra.

En el **ANEXO 5** se encuentra el diccionario de datos de la información recolectada y un análisis descriptivo de los archivos proporcionados por el CIAT y FENALCE.

3.3 PREPARACIÓN DE DATOS

El CIAT y FENALCE proporcionaron 8 archivos en los cuales se volcaron registros de la base de datos que contenían la información asociada a los cultivos de maíz, dichos ficheros fueron analizados para tomar la decisión de que datos se conservaban, removían o transformaban para definir con ello la vista minable.

3.3.1 Selección de archivos

El archivo Production_events_3.csv se encontraba filtrado, conteniendo únicamente registros pertenecientes al departamento de Córdoba, este fichero se conservó debido a la información contenida y su relación con los archivos restantes (ver archivos en fondo azul en la Figura 1). Por cada existencia de un registro al interior de este archivo se podían encontrar cero o muchos registros asociados en los demás documentos proporcionados por el CIAT y FENALCE. Inicialmente en Production_events_3.csv se encontraban 998 registros sobre cultivos de maíz, pero con la finalidad de obtener una variable objetivo uniforme, se conservaron solo 884 observaciones, en estas el producto cosechado fue grano seco, las restantes 114 presentaban otras opciones como: choclo y ensilaje (Con hojas y tallo); al tener una mayoría de registros con grano seco, se concentró el trabajo en este último tipo de cosecha y se descartó el resto.

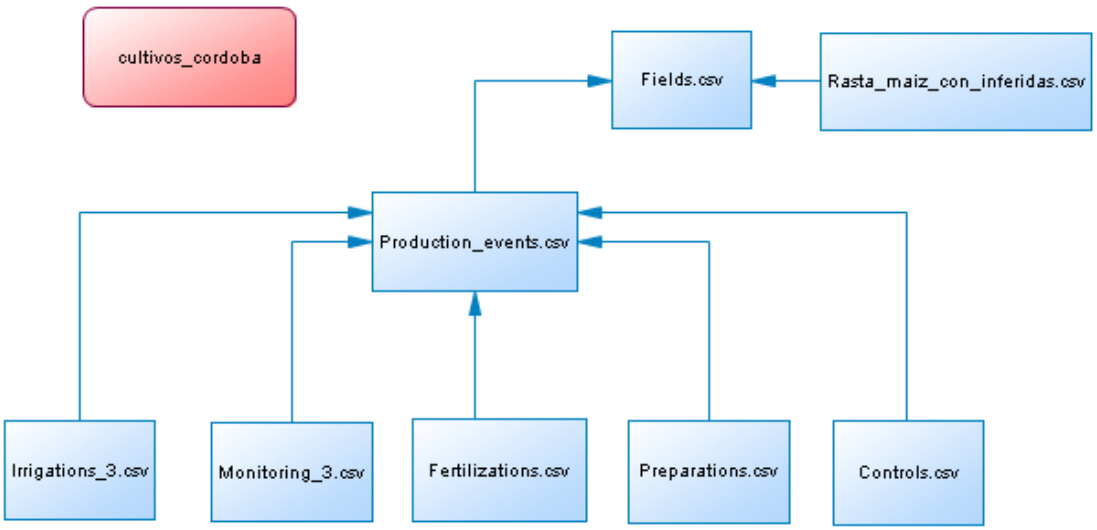


Figura 1. Representación de relación entre archivos proporcionados

El archivo rasta_maiz_con_inferida.csv se preservó, debido a que contenía 48 variables que aportaban conocimiento sobre características de 806 lotes que fueron

utilizados en los cultivos realizados al interior del departamento de Córdoba. Al hacer un **join** de las observaciones de rasta_maiz_con_inferida.csv con las 884 que se conservaron en Production_events.csv se obtuvieron 806 registros que estaban asociados en ambos documentos, los restantes fueron descartados debido a que al aplicar un proceso de imputación o llenado de datos en 48 columnas para 78 observaciones influía negativamente en los algoritmos que se utilizarían posteriormente en la obtención del modelo de clusterwise. La unión de estos 2 ficheros se denominó cultivos_cordoba.

El archivo Irrigations_3.csv contenía datos sobre los riegos realizados a los cultivos, pero no contenía registros pertenecientes al departamento de Córdoba. La razón de esta falta de información según lo informado por investigadores del CIAT, fue debido a que en este departamento no se presentaba a la fecha de la recolección de la información una implantación de esta práctica agrícola. Por lo tanto, al generar un aporte nulo en conocimiento, este archivo fue descartado.

El archivo Monitoring_3.csv contenía datos sobre monitoreos realizados a los cultivos. Este fichero presentaba 608 registros pertenecientes al departamento de Córdoba y al validar la existencia de una asociación de estas observaciones con las incluidas en cultivos_cordoba se obtuvo que de los 806 registros 530 se quedaron sin registros de monitoreo asociados. Esto indicó que la práctica agrícola de monitoreo no fue ampliamente desarrollada en los cultivos y debido al bajo aporte, este archivo fue descartado.

El archivo Controls_3.csv contenía datos sobre controles realizados de malezas, plagas y enfermedades, este fichero presentaba 3128 registros pertenecientes al departamento de Córdoba. Después de validar la existencia de una asociación de estas observaciones con las incluidas en cultivos_cordoba se obtuvo que de los 806 registros solo 7 registros quedaron sin controles asociados. Por lo anterior, el archivo se conservó y su información se tuvo en cuenta para la generación de la vista minable.

El archivo Fertilizations_3_Maiz_limpia.csv contenía datos sobre fertilizaciones aplicadas a los cultivos. Este fichero presentaba 2270 registros pertenecientes al departamento de Córdoba y luego de validar la existencia de una asociación de estas observaciones con las incluidas en cultivos_cordoba, se obtuvo que de los 806 registros solo 52 registros quedaron sin fertilizaciones asociados. Por lo anterior, el archivo se conservó y su información se tuvo en cuenta para la generación de la vista minable.

El archivo Preparations_3.csv contenía datos sobre la realización de preparativos previos a la siembra de los cultivos. Este fichero presentaba 612 registros pertenecientes al departamento de Córdoba y después de validar la existencia de una asociación de estas observaciones con las incluidas en cultivos_cordoba se obtuvo que, de los 806 registros, 447 se quedaron sin valores de preparación asociados. Esta cantidad de registros faltantes permitían interpretar que la práctica preparación del terreno y su información no fue completamente registrada o quizás

no se realizó. Debido al poco aporte de la información en este archivo, este fue descartado.

El archivo Fields_3.csv contenía datos sobre lotes registrados para ser utilizados en siembras. Este fichero presentaba 4004 registros pertenecientes al departamento de Córdoba y luego de validar la existencia de una asociación de estas observaciones con las incluidas en cultivos_cordoba se obtuvo que los 806 registros contaban con un respectivo lote asociado. Por lo anterior, el archivo se conservó y su información se tuvo en cuenta para la generación de la vista minable.

De las 806 observaciones que se conservaron, solo 799 contaban con información de clima. A los registros restantes no fue viable realizarles el procesamiento por parte del CIAT, esto debido a que no se justificaba el esfuerzo para un número tan pequeño de observaciones, por esta razón se descartaron estas 7 observaciones.

Tabla 2. Ubicación geográfica de los lotes caracterizados con RASTA

Municipio	***Total de Lotes***	***Coordenadas***
Cereté	173	8°53′12″N 75°47′28″O
Chimá	69	9°08′56″N 75°37′44″O
Ciénaga de Oro	84	8°52′30″N 75°37′16″O
Cotorra	171	9°02′20″N 75°47′36″O
Lorica	134	9°14′19″N 75°48′50″O
Montería	37	8°45′35″N 75°53′08″O
Purísima de la Concepción	6	9°14′11″N 75°43′25″O
San Carlos	50	8°47′40″N 75°41′58″O
San Pelayo	66	8°57′28″N 75°50′15″O
Tierralta	2	8°10′22″N 76°03′34″O
Valencia	14	8°15′33″N 76°08′49″O

Figura 2. Mapa político del departamento de Córdoba

Tabla 3. Tipo de información aportada por archivo.

Archivo	*Manejo Agrícola*	*Características del Suelo*	*Información Climática*	*Situación*
Production_events.csv	**X**		**X**	**Conservado**
rasta_maiz_con_inferida.csv		**X**		**Conservado**
Irrigations_3.csv	**X**			**Descartado**
Monitoring_3.csv	**X**			**Descartado**
Controls_3.csv	**X**			**Conservado**
Fertilizations_3_Maiz_limpia.csv	**X**			**Conservado**
Preparations_3.csv	**X**			**Descartado**
Fields_3.csv		**X**		**Conservado**

3.3.2 Tratamiento de la información

Una vez se definieron los archivos a ser conservados por aportar información, se continuó con la toma de decisiones sobre la preservación, tratamiento y transformación de las variables contenidas en estos, con la finalidad de generar un conjunto de datos sin información faltante, redundante, errónea, o de nulo aporte en términos de conocimiento para los algoritmos a ser utilizados en la adaptación del enfoque clusterwise y la posterior optimización.

Todas las variables contenidas en los archivos conservados fueron tratadas con técnicas de pre procesamiento de información estándar en proyectos de minería de datos como, por ejemplo, la detección de valores atípicos basado en rango Inter cuartil, análisis de conteos por clase, entre otros. Pero es preciso destacar que existieron escenarios en donde se debió dar un procesamiento especial a algunos datos, a continuación, se exponen estos casos.

- **Ajuste a la variable de rendimiento**: Esta variable tenía como unidad de medida kg/h en relación con el grano seco cosechado. Los registros almacenados en esta columna fueron transformados para obtener un valor más aproximado al real, esto debido a que la humedad influía en el peso final y los diferentes registros indicaron que los granos no presentaban el mismo porcentaje de humedad al momento de la cosecha. Investigadores del CIAT establecieron 14% como el valor promedio de humedad recomendado para el grano al momento de la cosecha. Para obtener los nuevos valores se utilizaron las variables de: rendimiento inicialmente registrado, la humedad contenida por el grano al momento de ser recolectado y la humedad recomendada, con estas variables se construyó la **Ecuación** 1 para definir el nuevo peso en kg/h.

$$RDT\_AJUSTADO = \frac{100 - \text{HUMEDAD}}{100 - 14} * \text{RENDIMIENTO} \qquad 1$$

- **Ajuste en las fechas asociadas a las etapas de cultivo**: En Production_events_3.csv se realizó un ajuste a 11 registros (con id 293, 830, 2017, 4620, 2302, 2510, 2441, 2292, 3178, 4513 y 4659), los cuales presentaban incoherencias en alguna de las fechas asociadas a las etapas del cultivo (siembra, emergencia, florecimiento y cosecha). El ajuste se desarrolló partiendo del valor de la media que se tenía para cada lapso de tiempo transcurrido en días entre las etapas, además también se tuvo presente el valor de la media que daba la duración en días de los cultivos (fecha cosecha – fecha emergencia de plantas), así, entonces aquellas fechas identificadas como incorrectas fueron manualmente cambiadas para que encajaran con el valor de la media al calcular los días trascurridos entre las etapas continuas, y al actualizar este valor se revisó que no se afectara el número de días del cultivo con respecto a su media, en caso de ocurrir esta situación se procedía a revisar las fechas restantes comparando con las respectivas medias y observar donde se debía realizar el pertinente ajuste de fechas.
- **Ajuste a columnas con valores dependientes**: En algunas columnas se asignó un valor especial, esto debido a la identificación de casos en donde existía dependencia entre las variables. Esta situación se presentó en los registros relacionados con la guía RASTA. Se observó dependencia entre la identificación de piedras, rocas, raíces y otras con la profundidad en la que se encontraban estas. Cuando había ausencia de datos en la primera columna en las dependientes no se registraba valor alguno, esto ocasionaba que la columna aparentara tener información faltante, lo cual no era correcto, por esto fue necesario identificar en qué casos si debía existir un valor en lugar de un espacio en blanco o valor nulo, permitiendo conocer con ello que era un caso especial y no una ausencia del dato. Se escogió el valor -1 para hacer esa distinción, debido a que era imposible en todos los casos que se tomara este como un valor real; porque las columnas dependientes almacenaban la profundidad (valores positivos) a la que se identificaba las piedras, rocas, raíces, etc.
- **División de la variable en valores porcentuales:** Algunos datos recolectados siguiendo la guía RASTA fueron reportados de una manera particular como se puede observar en la **Tabla 4**. Este es el caso de las variables que representaban el color de la tierra estando húmeda y seca, además su textura, resistencia al rompimiento y material orgánico. En el ejemplo de la **Tabla 4** las observaciones en la columna TEXTURA variaban en la cantidad de valores y era debido al número de capas que se observaron al abrir el cajón en la tierra como lo sugería la guía. Al no contar con la misma cantidad de observaciones y variando en importancia debido a que la presencia de estos valores estaba ligado al grosor en cm de cada capa encontrada, se tomó la decisión de desarrollar un proceso de **división de la variable TEXTURA en varias columnas por valores porcentuales (division of variable into several columns based on percentage values)**. Este proceso consistió en crear *N*

nuevas columnas y cada una de estas *N* columnas representarían uno de los posibles valores que podría existir al interior de la variable, ahora entendiendo que la profundidad del cajón representaba el máximo valor de grosor posible que podría tomar una capa, se tomó esto como el 100%, entonces para cada uno de los valores reportados junto con su grosor observado se calculó en términos de porcentaje la incidencia de esta observación al interior del cajón y se almacenó en su respectiva columna como se puede observar en la **Tabla 5**. Esta representación tiene el inconveniente de perder información relacionada con el orden de ocurrencia de las capas (desde la más superficial a la más profunda), pero para mantener el alcance del proyecto en control, no se consideró, esta decisión se socializo con expertos del CIAT y se comparó con otra forma que tiene en cuenta la capa de mayor espesor identificada dentro de la profundidad efectiva del maíz, concluyendo que se logra guardar mayor información con la división propuesta sin importar el orden de ocurrencia.

Tabla 4. Representacion inicial de variables asociadas a textura

ID_LOTE	NO_CAPAS_RASTA	ESPESOR	TEXTURA
40	3	22,9,50	FAr, FrL, FA
43	3	20,14,36	FAr, FAr, ArL
44	3	26,14,38	FAr, FAr, AF
45	3	16,39,29	FAr, FrL, FrL
46	3	33,6,47	FAr, FAr, FrL
47	3	10,20,24	FAr, FAr, FrL
51	3	18,33,30	FAr, ArA, Ar

Otras variables que surgen de la aplicación de la guía RASTA corresponden a la color de la tierra húmeda y seca, los posibles valores que pueden tomar son 54 y no fue posible disminuir este número de valores, por lo que si se generaba una nueva columna por cada posible valor, esto implicaría tener que trabajar con una dimensionalidad más alta de lo que ya se tenía para los 799 registros de datos existentes, por lo tanto, se decidió excluir estas variables de la vista minable. Además los expertos del CIAT consideran que debido al proceso de generación de variables inferidas, las variables que representaban el color se derivan en otras como: materia orgánica y drenajes, por tal razón esta información está contenida allí, finalmente se espera que al removerlas no se ocasione una pérdida significativa de información.

Tabla 5. Representacion final de variables asociadas a textura

ID_LOTE	Porc_A	Porc_Ar	Porc_ArA	Porc_ArL	Porc_FrL	Porc_L	Porc_F	Porc_FAr	Porc_FA	Porc_Af
40	0	0	0	0	11.11	0	0	27.16	61.73	0
43	0	0	0	51.43	0	0	0	48.57	0	0
44	0	0	0	0	0	0	0	51.28	0	48.72
45	0	0	0	0	80.95	0	0	19.05	0	0
46	0	0	0	0	54.65	0	0	45.35	0	0
47	0	0	0	0	44.44	0	0	55.56	0	0
51	0	37.04	40.74	0	0	0	0	22.22	0	0

- **Transformación de fechas**: En cuanto a las fechas encontradas en Production_events_3.csv por si solas no brindaban mucha información para los algoritmos de minería de datos, por esto fueron utilizadas para generar conocimiento del número de días transcurridos entre las etapas del cultivo: siembra – emergencia, emergencia – floración y floración – cosecha, y así brindar conocimiento para estos tipos de algoritmos.

  Por otra parte, las fechas también fueron utilizadas en conjunto con los datos obtenidos sobre fertilizaciones y controles; para esto se identificaron los tipos de fertilizaciones y controles realizados, con lo que se procedió a contar las repeticiones de estos tipos de eventos dentro del tiempo transcurrido entre las etapas del cultivo, para las fertilizaciones de manera adicional se sumó el contenido de nitrógeno, potasio y fosforo aplicado. Con este proceso se integró a la vista minable los datos pertenecientes a los archivos Fertilizations_3_Maiz_limpia.csv y Controls_3.csv con Production_events_3.csv.

Con los datos integrados en un único archivo se procedió a realizar un análisis exploratorio sobre los registros contenidos, eliminando variables en las que se observó 0 variación o que fueron consideradas como identificadores y aquellas con más de un 20% de valores faltantes, también se identificaron y eliminaron registros duplicados y erróneos que no se lograron ajustar.

Teniendo ya un conjunto de registros unificados y limpios pero aún con datos faltantes, fue necesario completar estas observaciones, lo primero que se trató de hacer, fue recuperar los datos desde las fuentes (agricultor, FENALCE y CIAT), pero esta opción no fue viable debido a los tiempos que emplearía realizar el contacto y la poca garantía de recuperar la información, por lo que se utilizaron dos técnicas; la primera consistió en que si el total de observaciones faltantes en la variable era menor al 1%, entonces estos valores se remplazaron por la media o moda dependiendo de la escala de la variable, la segunda era cuando el valor superaba al 1% y era menor al 20%, aquí se recurrió a la imputación de datos utilizando Random Forest debido a que sus predicciones son bastante precisas y puede manejar un número grande de variables de entrada [41], aunque el uso de KNN para esta actividad era otra opción, no se optó por este algoritmo debido a que no se contaba con un conjunto de observaciones que representaran la mayoría de las posibles combinaciones de eventos de cultivos de maíz al interior de Córdoba, con lo que los registros actuales no se encontraban en condiciones suficientemente similares para replicar algunos de sus valores dentro de otras observaciones que lo requirieran.

3.3.3 Reducción de variables

Con las variables finales de clima (27 variables), suelo (50 variables) y prácticas de manejo (37 variables) integradas en un único archivo, se contó con 799 registros y 115 variables. Para un archivo con este número de columnas era recomendable disminuir el número de características (variables), ya que de esta forma los algoritmos disminuirían su tiempo de procesamiento y además se debía buscar

eliminar características irrelevantes, ruidosas y altamente correlacionadas que afectaran los algoritmos.

Por todo anterior se analizaron 3 maneras de reducir la dimensionalidad, a saber: alta correlación, wrapper y baja varianza. A continuación, se describen cada uno de ellos.

3.3.3.1 Alta correlación

Al usar la herramienta RapidMiner y la función pairs.panels de R para detectar alta correlación (entre -1.0 y -0.8 o entre 0.8 y 1.0) entre las 115 variables, se obtuvieron indicadores sobre correlaciones altas como las de Temp_Max_Avg_Mad y Temp_Max_34_Freq_Mad, en donde no era adecuado definir cuál eliminar y cuál conservar sin tener el conocimiento o asesoría de un experto sobre la influencia de estas en los cultivos de maíz, conocimiento o asesoría que no se logró conseguir.

Además existieron casos donde había una alta correlación simultánea, como fue el caso que se observa en la **Figura 3,** en donde el grupo de variables Temp_Max_Avg_Veg, Temp_Max_34_Freq_Veg y Temp_Avg_Veg, presenta entre las 2 primeras una correlación del 93% y para las 2 últimas del 83% y finalmente para la primera con la última de un 90%. En estos casos tampoco fue posible tomar una decisión sobre cuáles variable conservar y cuáles descartar sin un buen conocimiento en el área, conocimiento que no se logró conseguir.

First Attribute	Second Attribute	Correlation ↓
TotN_Antes_Siem	FerQui_Antes_Siem	0.975
Temp_Max_Avg_Mad	Temp_Max_34_Freq_Mad	0.956
Temp_Max_Avg_For	Temp_Max_34_Freq_For	0.952
Rain_Accu_Veg	Rain_10_Freq_Veg	0.942
Temp_Max_Avg_Veg	Temp_Max_34_Freq_Veg	0.928
Temp_Max_Avg_Veg	Temp_Avg_Veg	0.899
Rain_Accu_For	Rain_10_Freq_For	0.885
PENDIENTE_RASTA	drenaje_externo	0.872
Temp_Max_Avg_Veg	Diurnal_Range_Avg_Veg	0.861
Temp_Max_Avg_For	Temp_Avg_For	0.861
Temp_Min_Avg_Mad	Temp_Avg_Mad	0.854
Rain_Accu_Mad	Rain_10_Freq_Mad	0.829
Temp_Max_Avg_Mad	Temp_Avg_Mad	0.829
Temp_Min_Avg_For	Temp_Avg_For	0.827
Temp_Avg_Veg	Temp_Max_34_Freq_Veg	0.825

Figura 3. Correlación entre variables

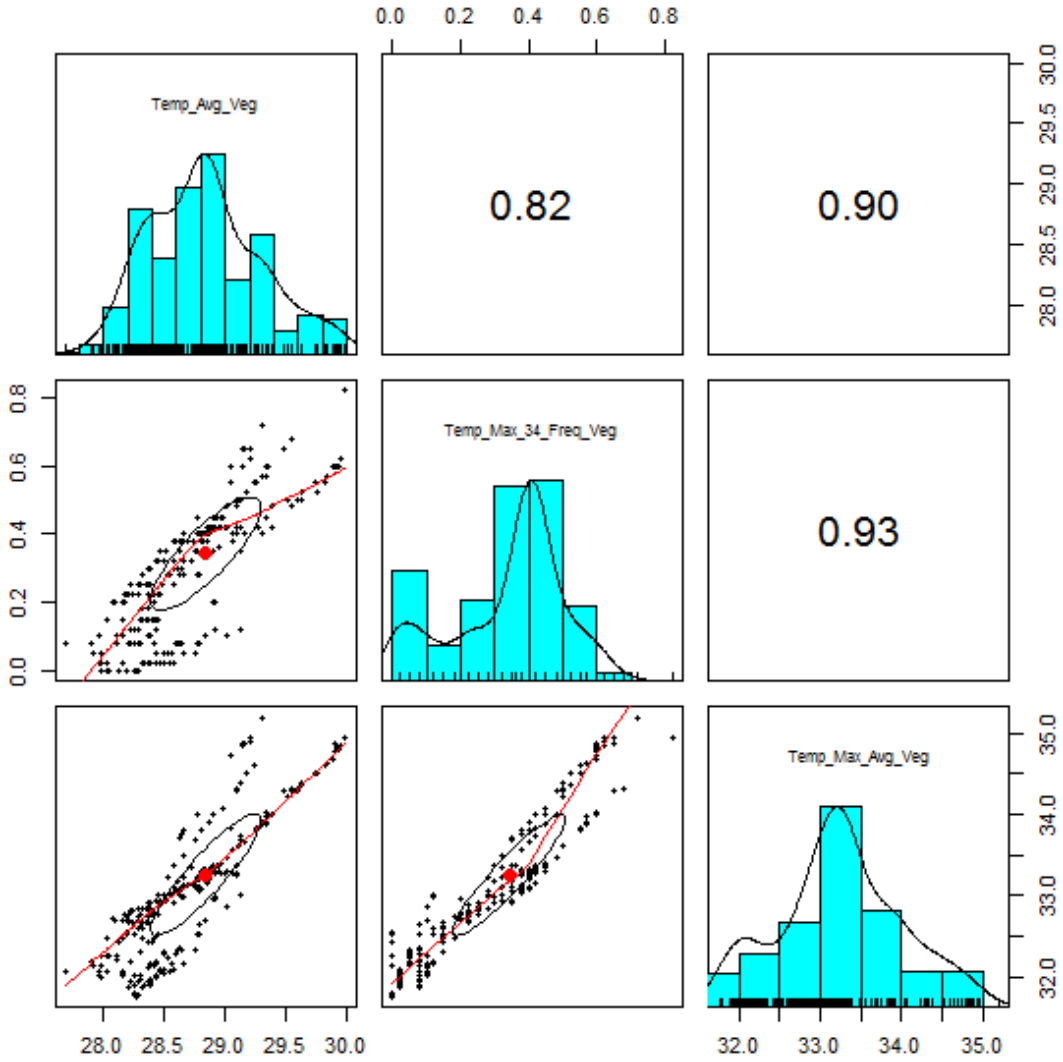


Figura 4. Diagrama de dispersión y coeficiente de correlación

Por lo anterior, la opción de reducir el número de variables partiendo de la correlación existente entre ellas se dejó como una actividad a desarrollar por parte de Weka con el clasificador de LinearRegression habilitando la opción de eliminateColinearAttributes.

3.3.3.2 Wrapper

Se usó el WrapperSubsetEval con el método de búsqueda GreedyStepwise de Weka para hacer la selección de atributos (**ANEXO 7**). Este algoritmo parte de un conjunto de datos como entrada, con estos registros procede a generar múltiples modelos de regresión y en cada una de estas se evalúa el nivel de calidad, al final entrega en un listado las variables que considera como influyentes.

El wrapper fue ejecutado 37 veces, en cada una de estas el conjunto de datos entrante era el mismo y las listas obtenidas de variables seleccionadas como influyente fueron distintas, con estas 37 listas se identificó el número de apariciones por cada variable, se observó que existían variables que fueron seleccionadas en la totalidad de las ejecuciones y algunas que nunca lo fueron.

Una vez identificado el número de ocasiones en que se seleccionaron las variables, se ordenaron de mayor a menor según la frecuencia de aparición; con lo que las seleccionadas como relevantes en más ocasiones quedaron de primeras. Con los datos ordenados se calculó el número de apariciones en términos de porcentaje, luego se generó una nueva columna donde se almacenó un porcentaje acumulado.

La columna con los porcentajes acumulados se utilizó para obtener una interpretación de hasta donde un grupo de variables aportaban importancia o conocimiento al problema, por ejemplo: al seleccionar las primeras 5 variables se observó que estas tenían un porcentaje acumulado de 12.99 como se ve en la **Tabla 6**. Partiendo de esto se obtuvieron dos nuevas vistas minables, una contenía las variables hasta donde se obtenía un porcentaje acumulado del 90% y la segunda hasta donde el porcentaje acumulado llegaba al 95%.

Tabla 6. Incidencias en la selección de variables por Wrapper

Variable	Apariciones	Porcentaje	Acumulado	Porcentaje Acumulado
METODO_COSECHA	37	100.0	100.0	3.248463565
TIPO_MATERIAL	37	100.0	200.0	6.496927129
TotP_Siem_Emer	37	100.0	300.0	9.745390694
OBSERVA_MOHO_RASTA	37	100.0	400.0	12.99385426
PROFUND_RAICES_VIVAS_RASTA	37	100.0	500.0	16.24231782
OBSERVA_PLANTAS_PEQUENAS_RASTA	37	100.0	600.0	19.49078139
Temp_Min_Avg_Mad	37	100.0	700.0	22.73924495
MATERIAL_GENETICO	36	97.3	797.3	25.8999122
Sol_Ener_Accu_Veg	36	97.3	894.6	29.06057946
REGION_SECA_ARIDA_RASTA	35	94.6	989.2	32.1334504
ALTURA_LOT	34	91.9	1081.1	35.11852502
Rain_Accu_For	34	91.9	1173.0	38.10359965
Diurnal_Range_Avg_Veg	33	89.2	1262.2	41.00087796
ContMalMec_Flor_Cose	31	83.8	1345.9	43.72256365
Rain_10_Freq_For	31	83.8	1429.7	46.44424934
FerOrg_Emer_Flor	30	81.1	1510.8	49.07813872
OBSERVA_COSTRAS_BLANCAS_RASTA	29	78,4	1589.2	51.62423178
CULT_ANT	28	75.7	1664.9	54.08252853
Sol_Ener_Accu_Mad	25	67.6	1732.4	56.27743635
PROFUND_MOTEADOS_RASTA	24	64.9	1797.3	58.38454785
Rhum_Avg_For	23	62.2	1859.5	60.40386304

Posteriormente, al ejecutar las metaheurísticas con enfoque de clusterwise usando el conjunto de datos sin aplicar reducción de variables de ningún tipo, se identificó que se obtenían mejores valores de R^2 *ajustados* que los obtenidos con los archivos que contenían sólo el 90% y 95% de las variables. Por esta razón esta estrategia también fue descartada.

3.3.3.3 Baja varianza

Partiendo de que se conocía que el enfoque clusterwise iba a dividir el conjunto de datos de la vista minable en *N* grupos, se pensó que, si una variable presentaba una baja variación en el conjunto de datos total, entonces al dividir estos en los grupos se iba a presentar una variación aún más baja o incluso de cero en la variable en cuestión al interior de los grupos resultantes.

Entendiendo que el algoritmo de regresión lineal de Weka remueve variables con varianza cero, se procedió a eliminar columnas con esta característica; es decir aquellas donde todos sus registros tomaban el mismo valor, además se eliminaban las que presentaban un conjunto pequeño de valores y la proporción de la frecuencia del valor más común frente a la frecuencia del segundo valor era grande, es decir, donde existía una presencia dominante de cierto valor.

Al ejecutar los experimentos del enfoque clusterwise con el conjunto de datos sin aplicar reducción de variables de ningún tipo se identificó que algunas de las variables que eran candidatas a ser eliminadas por su baja varianza si estaban quedando seleccionadas en los modelos de regresión para algunos clústeres, lo que indicó que dentro de estos grupos los valores eran influyentes y además al partir el conjunto de datos la varianza podía aumentar en algunos casos. Por esta razón esta estrategia también fue descartada.

Una vez descartadas las 3 opciones para reducir la dimensionalidad del conjunto de datos, se optó por trabajar como vista minable la opción inicial de 115 variables y 799 observaciones y dejar la tarea de reducir la dimensionalidad a Weka cuando realice los modelos de regresión lineal de cada clúster, esto se explica en el siguiente capítulo.

CAPÍTULO 4

4 ADAPTACIÓN DEL ENFOQUE CLUSTERWISE A CULTIVOS DE MAIZ

Este capítulo presenta la adaptación del enfoque clusterwise usando las metaheurísticas MSSA y GRASP; mediante una descripción del proceso realizado en el cual también se incluye la comparación y conclusión de los resultados obtenidos.

Partiendo de la vista minable previamente definida se realiza el proceso de clusterwise, iniciando con un proceso denominado “barajado inicial” y luego con un segundo proceso denominado “optimización” donde finalmente se obtienen los *K* grupos o clústeres. Cada grupo contiene un conjunto de elementos (registros o tuplas) que se ajustan mejor a un modelo de regresión, ajuste que es medido por el valor R^2 *ajustado* partiendo de que los modelos obtenidos son estadísticamente significativos.

El **barajado inicial** es el proceso encargado de distribuir las observaciones en diferentes grupos para generar un punto de partida al proceso de optimización. Para esto se cuenta con un número *K* de grupos (clústeres) previamente establecido. Cada grupo tiene asociado un conjunto de *N* observaciones similares bajo el criterio de distancia utilizado y se representa como se muestra en la **Figura 5**.

K1							K2							…	K5						
1	10	799	…	11	51		4	720	8	…	40	32			2	12	622	…	7	614	

Figura 5. Representación de agrupamiento (ejemplo de 5 clústeres)

Para el proceso de barajado inicial se optó por utilizar el algoritmo k-means, usando el criterio de distancia euclidiana para variables mixtas, esto teniendo en cuenta que los datos de cultivos de maíz cuentan con variables binarias, continuas y nominales.

Posterior al barajado inicial el resultado de distribución lo usan las metaheurísticas MSSA o GRASP en el proceso de **optimización** sobre el espacio de búsqueda en un periodo de tiempo determinado. El proceso de búsqueda se apoyó en la regresión lineal múltiple, la cual se usó para generar modelos de regresión a cada grupo de observaciones, tomando como variable dependiente el rendimiento de los cultivos.

Los modelos generados para cada clúster tienen asociado un R^2 *ajustado*, luego, se calcula la media (promedio) de los R^2 *ajustados* conseguidos en los *K* modelos y el valor obtenido se toma como medida de calidad de la solución (*fitness*).

En el proceso de búsqueda de nuevas soluciones se utilizan dos criterios de factibilidad con el cual se define si una solución es o no válida, el primero es que para cada grupo se haya podido crear su respectivo modelo de regresión y segundo,

que cada uno de estos modelos tenga el valor de significancia estadística (*valor P*) por debajo de 0.05 con lo que se garantiza que entre las variables independientes y la dependiente existe realmente una relación lineal significativa.

Finalmente, del proceso de búsqueda de una solución óptima se logra obtener un macro modelo conformado por un conjunto de *K* modelos de regresión lineal y sus *K* correspondientes centroides, los cuales están asociados a cada grupo encontrado en la mejor solución. Estos modelos de regresión lineal múltiple describen e identifican un conjunto de variables que tienen un nivel de importancia e inciden sobre su respectivo grupo de observaciones (cultivos) las cuales son muy similares bajo el criterio de R^2 *ajustado* asociado a cada modelo.

Con los modelos encontrados se apoya la toma de decisiones, teniendo en cuenta que uno de los principales problemas en la agricultura es identificar prácticas de manejo que generen efectos positivos sobre los cultivos. Con el uso del enfoque clusterwise se pueden determinar estas prácticas de manejo basadas en agricultura específica por sitio donde se identifican prácticas que inciden sobre el rendimiento de un cultivo y estas surgen de datos de otros cultivos con condiciones similares.

4.1 MEDIDAS DE CALIDAD

4.1.1.1 Coeficiente de determinación ajustado

El coeficiente de determinación ajustado (R^2 *ajustado*) es una medida de calidad o bondad de ajuste a un modelo lineal. Este determina el porcentaje de variación de la variable dependiente con respecto a las variables independientes incluidas en un modelo, a diferencia del coeficiente de determinación (R^2) este es menos sensible al número de variables incluidas en un modelo si no son significativas. La **Ecuación 2** presenta la fórmula de R^2 *ajustado*.

$$R^2_{Ajustado} = 1 - \frac{\mathrm{N} - 1}{\mathrm{N} - \mathrm{nv} - 1} * (1 - R^2) \qquad 2$$

Donde ***N*** es el número de observaciones que pertenece a un clúster y ***nv*** es el número de variables seleccionadas en el modelo para el clúster.

Debido a que el R^2 *ajustado* determina la calidad de un modelo de regresión lineal para un clúster, este también se utilizó para calcular la calidad de una solución del enfoque clusterwise que se considera como un macro modelo conformado por un conjunto de *K* modelos, para ello se definió la calidad de una solución (*fitness*) como el promedio de estos valores (ver **Ecuación 3**)**.**

$$fitness = \frac{\sum_{i=1}^{k} R^2_{Ajustado\ i}}{\mathrm{k}} \qquad 3$$

Suponiendo que se cuenta con una solución de 3 clústeres, es decir un $k = 3$, donde los modelos generados para cada uno de estos clústeres tienen un valor de R^2 *ajustado* correspondiente a 0.87, 0.85 y 0.75, El *fitness* para esta solución es de 0.82, el cual corresponde a la media de los valores de R^2 *ajustado* de los 3 modelos asociados a la solución.

4.1.1.2 Validación del modelo

La construcción de una solución usando el enfoque clusterwise implica tener *K* modelos asociados a un conjunto de *K* clústeres, por tanto, todos los modelos deben validarse de la siguiente forma:

- Cada grupo (clúster) debe tener un modelo de regresión lineal múltiple asociado. Aunque parece obvio, puede suceder que en el proceso de barajado inicial o en el proceso de optimización con MSSA o GRASP, al momento de distribuir las observaciones no siempre el resultado permita generar un modelo. Esto puede ser causado debido a una equivocada agrupación de observaciones donde estas no comparten las características que influyen sobre la variable dependiente para ese clúster, o porque en el clúster queda un número inferior de observaciones al de columnas.

- El *valor P* del modelo debe ser inferior al valor de significancia establecido en 0.05, con lo que se rechaza la hipótesis nula.

4.2 REPRESENTACIÓN DE LA SOLUCIÓN

Una solución del enfoque clusterwise es un macro modelo compuesto por *K* grupos (clústeres), cada uno de los cuales se conforma por un conjunto de observaciones (únicamente los ids de las filas en la vista minable), un modelo de regresión lineal múltiple asociado, el valor R^2 *ajustado* del modelo de regresión y el *valor P* de dicho modelo. La solución además de la información de los *K* grupos incluye el *fitness* (promedio de los valores de R^2 *ajustado* de los *K* grupos según la **Ecuación 3**) y se puede apreciar en la **Figura 6**. Esta representación se usa de manera trasversal tanto en el proceso de barajado inicial, como en el proceso de optimización del rendimiento de los cultivos de maíz que se presenta más adelante.

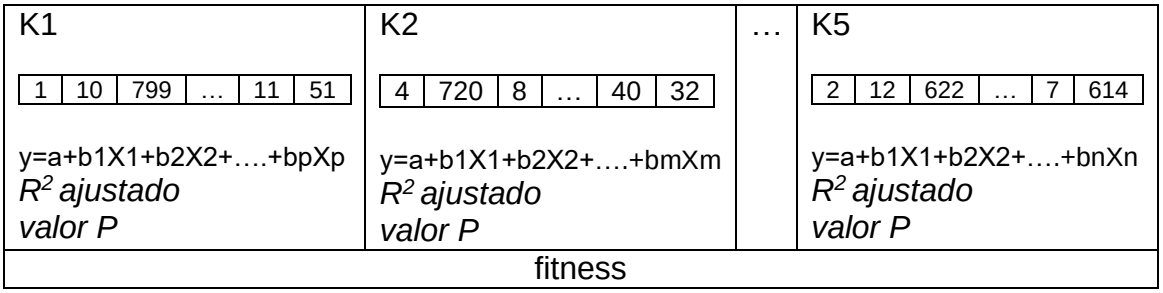


Figura 6. Representación de una solución en el enfoque clusterwise

Construcción de la solución

La construcción de cada modelo de regresión lineal se hizo usando la librería Weka 3.8. Para el uso de la librería se implementó un adaptador que se encarga de preparar los datos asociados a cada clúster y configurar los parámetros necesarios para ejecutar un proceso de clasificación usando la función LinearRegression (**ANEXO 8 - DataAdapter.java**).

El proceso de preparación de los datos se realizó para que Weka tuviera en cuenta los diferentes tipos de variables (continuas, binarias y nominales) de acuerdo con el problema, también definir la variable de clase, en este caso el rendimiento ajustado del cultivo (RDT_AJUSTADO) y además tener en cuenta cuales variables ignorar del dataset como el identificador (ID_LOTE). Algo muy importante del proceso fue la definición de las variables, debido a que Weka necesita declarar los tipos de variables y algunos datos adicionales como por ejemplo, los posibles valores que van a tomar estas (nominales y/o binarias), teniendo en cuenta esto el dataset completo con sus 799 observaciones tiene asociado una descripción para cada variable como se aprecia en la **Figura 7**. Pero debido a que cada clúster no tiene todas las observaciones estas descripciones se calculan por cada variable del centroide. Por ejemplo, en el dataset la variable material genético cuenta con 26 posibles valores de los cuales en un grupo de acuerdo con las observaciones tiene solo 8 de los 26 posibles, en ese sentido la estructura de control para esta variable solo registra esos valores y luego esta información la usa Weka para creación del modelo de regresión de ese grupo con los datos correctos.

Nombre		
Tipo		
Binaria	*Continua*	*Nominal*
Resumen		
Valor 1 = Conteo Valor 2 = Conteo **(Solo dos valores)**	Min = # 1st Qu. = # Mediana = # Media = # 3rd Qu. = # Max	Valor 1 = Conteo Valor 2 = Conteo …. Valor N = Conteo

Figura 7. Estructura de descripción de variable por tipo

Los parámetros utilizados en Weka para la construcción de los modelos fueron los que están definidos por defecto (ver la **Figura 8**) debido a que dan muy buenos resultados, adicionalmente se habilitó el uso de estadísticas adicionales y en las opciones de prueba se optó por usar el mismo dataset de entrenamiento (Use training set), en este caso son las observaciones asociados a cada clúster.

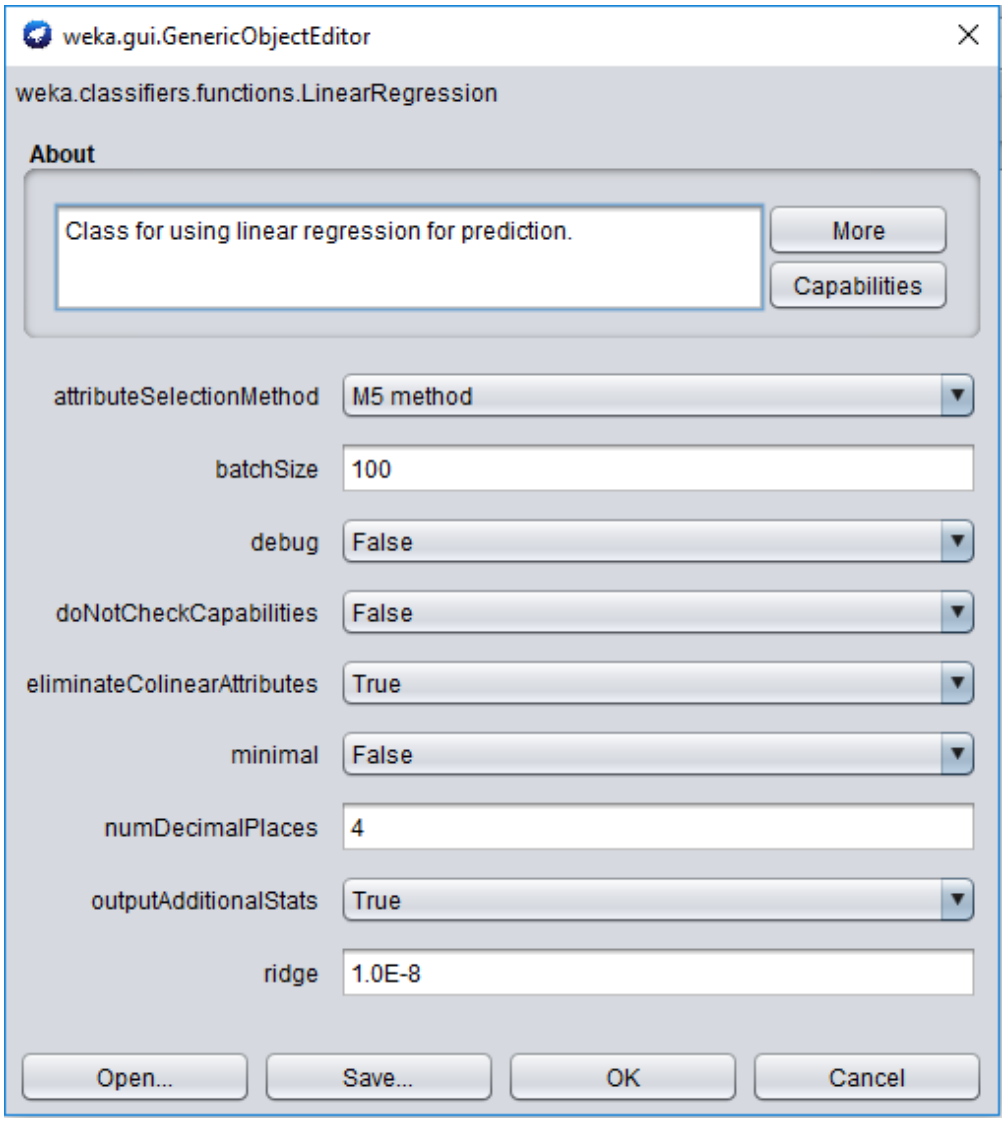


Figura 8. Parámetros de herramienta Weka

A continuación, se describen los diferentes parámetros utilizados para la construcción de los modelos de regresión lineal múltiple:

- AttributeSelectionMethod: indica el método de selección de atributos. El método elegido es M5, el cual construye un árbol de modelos donde cada hoja tiene asociado un modelo de regresión lineal múltiple. La construcción se basa en la eliminación de atributos y la disminución del error de predicción de cada modelo usando el criterio de información Akaike (Akaike Information Criterion, AIC).
- BatchSize: indica el tamaño del lote usado para la predicción, es decir, el número ideal de observaciones para la predicción por lotes cuyo valor por defecto es 100.
- doNotCheckCapabilities: cambia el estado de verificación de las capacidades del clasificador, cuyo valor por defecto es false (sin verificar), este parámetro en caso de colocar verdadero incrementa el tiempo de ejecución.

- EliminateColinearAttributes: habilita o deshabilita la eliminación de variables altamente correlacionadas, cuyo valor por defecto es habilitado (true).
- Minimal: sirve para habilitar o deshabilitar que conserve en memoria los diferentes valores de medias, desviación e información del modelo de regresión. Su valor por defecto es deshabilitado (false).
- NumDecimalPlaces: indica el número de cifras decimales en los resultados. Su valor por defecto son 4 cifras decimales.
- OutputAdditionalStats, sirve para habilitar o deshabilitar el cálculo de valores estadísticos adicionales como coeficientes de desviación, estadístico *f* del modelo, entre otros. Se usó el valor habilitado (true) con el fin de poder obtener datos adicionales como el valor *p* entre otros.
- Ridge: determina el valor máximo de evaluaciones de la función objetivo, se utilizó el valor por defecto de 1,0E-8.

Para poder extraer los valores de R^2 ajustado y el *valor P* de los modelos generados por Weka fue necesario modificar el código fuente de algunas clases de dicha herramienta ya que no estaban incluidas en el reporte de estadísticos adicionales que entrega. En el **ANEXO 8** (LinearRegression.java) se pueden apreciar dichas modificaciones.

4.3 DISTANCIA EUCLIDIANA MIXTA

La distancia euclidiana mixta es la medida de distancia utilizada en el proceso de barajado inicial y en el proceso de optimización. El método usado para el cálculo de esta medida se presenta en el **Pseudocódigo 1**.

El cálculo de la distancia euclidiana mixta se encarga de determinar la distancia de un evento (observación o registro) al centroide de un clúster, la fórmula de la distancia euclidiana se presenta en la **Ecuación 4**. Para su cálculo se debe tener en cuenta que las diferencias entre los componentes, como lo muestra la ecuación cambian según el tipo de variable involucrada.

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad 4$$

Las variables continuas presentan un caso especial en el cálculo de su diferencia, debido a que existen casos en donde algunas variables no aplican para un evento y toman el valor de -1. Este valor de -1 es una etiqueta para no ser tenido en cuenta para algunos cálculos. En el caso de la distancia con este caso especial la diferencia calculada consiste en comparar el valor de la variable en la observación y en el centroide; si ambos son iguales a -1, la diferencia es 0, en caso contrario, donde este valor es tomado únicamente por la observación o por el centroide, la diferencia será de 1.

Para el caso normal en las variables continuas cuando la observación y el centroide toman valores dentro del rango de la variable se hace una resta entre el valor

normalizado de la variable del evento y el valor normalizado de la variable del centroide y el resultado se eleva al cuadrado, los valores normalizados se utilizan para que la distancia total no se vea afectada por los diferentes rangos que se manejan, esto quiere decir que los valores de las variables con rangos más grandes no tengan mayor peso sobre las variables con rangos más pequeños.

En las variables nominales y binarias la diferencia consiste en una comparación de los valores para la variable del evento y del centroide, si estos son iguales la diferencia es de 0, en caso contrario es de 1.

Finalmente, la distancia entre una observación y el centroide es la sumatoria de las diferencias de los cuadrados variable a variable y a esta sumatoria se le saca la raíz cuadrada como se presenta en la **Ecuación 4**. Un detalle muy importante es que el algoritmo ignora una lista de variables en el cálculo de la distancia, variables como por ejemplo el identificador de un evento (ID_LOTE) y la variable dependiente (RDT_AJUSTADO).

```
Entrada:    Centroide de un clúster centroide,
            Observación obs,
            Lista de variables a ignorar listIgno,
Salida:     Distancia entre el centroide y la observación
1   distancia = 0
2   si centroide.numeroVariables () == obs.numeroVariables () entonces
3       obsC = centroide.observacion
4       para i = 0 hasta obs.numeroVariables () - 1 hacer
5           si listIgno == nulo || listIgno.contiene (i) == falso entonces
6               diferencia = 0
7               varOC = obsC.obtenerVariable(i)
8               varO = obs.obtenerVariable(i)
9               si varOC es Continua && varO es Continua entonces
10                  si varOC.valor == -1 && varO == -1 entonces
11                      diferencia = 0
12                  si no
13                      si varOC.valor == -1 || varO == -1 entonces
14                          diferencia = 1
15                      si no
16                          diferencia = elevar((varOC.valorNormalizado –
                                varO.valorNormalizado) , 2)
17                      fin si
18                  fin si
19              si no
20                  si varOC es Nominal && varO es Nominal ||
                       varOC es Binaria && varO es Binaria entonces
21                      si varOC.valor == varO.valor entonces
22                          diferencia = 0
23                      si no
24                          diferencia = 1
25                      fin si
26                  fin si
27              fin si
28              distancia = distancia + diferencia
29          fin si
30      fin para
```

```
31  fin si
32  distancia = raíz (distancia)
33  retornar distancia
```

Pseudocódigo 1. Algoritmo para calcular la distancia euclidiana mixta

4.4 BARAJADO INICIAL

El proceso de barajado inicial hace uso del algoritmo k-means para realizar la distribución inicial de las 799 observaciones en *K* grupos, el funcionamiento de este algoritmo se presenta en el **Pseudocódigo 2.**

```
Entrada:    Numero de clústeres k,
            Numero de iteraciones n,
            Dataset data,
            Tipo de distancia tDis,
            Ignorar variables ignorar,
            Lista de variables a ignorar listIgno,
Salida:     Lista de k clústeres
1   clusters [] = initClusters (obtenerKCentroides (k, data))
2   mientras iteracion <= n hacer
3       para i = 0 hasta data.observaciones.tamaño - 1 hacer
4           obs = data.obtenerObservacion (i)
5           asignarClusterCercano (clusters, obs, tDis, ignorar, listIgno)
6       fin para
7       para i = 0 hasta k – 1
8           recalcularCentroide (clusters [i])
9       fin para
10      Iteración = iteración + 1
11  fin mientras
12  retornar clusters []
```

Pseudocódigo 2. Algoritmo K-means

Teniendo en cuenta que el algoritmo k-means está compuesto por 3 pasos principales como se describe en la **sección 2.1.6**, a continuación, se presentan los pasos con la respectiva adaptación al problema.

Paso 1: Se seleccionaron de forma aleatoria *K* observaciones del dataset, para asignar sus valores a los centroides iniciales de los *K* clústeres (Línea 1).

Pasó 2: Se asignan las 799 observaciones a los *K* clústeres, esta asignación se realiza bajo un criterio de distancia al clúster más cercano, es decir que se calcula la distancia de una observación al centroide de cada clúster y se elige el de menor distancia. Para medir la diferencia de longitud entre un centroide y una observación se elige la distancia euclidiana debido a que presento mejor desempeño en la etapa de experimentación (también se evaluó la distancia de Gower). Para la implementación se usó una lista de variables a ignorar dentro del dataset para que no se tuvieran en cuenta en el cálculo de distancias (ID_LOTE y RDT_AJUSTADO por ejemplo). Este paso corresponde a las líneas 3 a 6.

Paso 3: Se recalcula los centroides de los *K* clústeres, este procedimiento se hace teniendo en cuenta las observaciones que lo conforman. Un centroide tiene una

estructura similar a una observación estando compuesto por las 115 variables y una estructura de control por cada variable que aporta información adicional ayudando a disminuir cálculos durante su actualización (ver líneas 7 a 9).

La representación de un centroide es un vector de 115 variables (las mismas de la vista minable). Para las columnas de tipo continuas se guarda un promedio de los valores que tienen las observaciones, para las nominales y binarias se guarda el conteo de cada uno de los valores que tienen las observaciones. La **Figura 9** presenta un ejemplo de un centroide que corresponde a un clúster de 200 observaciones.

1	2	3	4	…	115
∑1120	Mecanizado 120 Manual 80	SI 99 NO 101	PIONEER 30F32 37 DK 234 55 FNC 114 48 Otro 60		∑1020120
V=5,6	V=Mecanizado	V=NO	V=Otro		V=5100.6

Figura 9. Representación de un centroide de un clúster

Teniendo en cuenta la **Figura 9**, las posiciones 1 y 115 de la figura pertenecen a una variable continua y almacenan la sumatoria de los valores que tienen las 200 observaciones en esas variables, las posiciones 2, 3 y 4 como son variables binarias y nominales tienen un conteo de los valores que tienen las 200 observaciones esto quiere decir que para la posición 2, existen 120 observaciones que tienen el valor mecanizado y 80 el valor manual. Con esta representación, cuando se agrega o retira una observación en un clúster es más rápido actualizar las estructuras de control de las 115 variables del centroide.

Para determinar el valor que el centroide usa como representante en cada una de las 115 variables se toma la media para las variables continuas y para las variables binarias o nominales se usa la moda. Para el ejemplo de la **Figura 9**, el valor que toma el centroide en la posición 2 es mecanizado porque corresponde a la moda en esa variable y para la posición 115 el valor a tomar es 5100.6 porque es la media de las 200 observaciones que pertenecen al clúster (1020120/200).

Finalmente, al terminar la ejecución del algoritmo k-means se obtiene un conjunto de *K* grupos de observaciones como se representa en la **Figura 5**; este resultado se valida utilizando la medida de calidad expuesta en la **sección 4.1.1.2** con el objetivo de asegurar que sea una solución viable (factible) para seguir con el proceso de optimización que se puede realizar con MSSA o GRASP.

Si el proceso de distribución inicial de observaciones en los *K* clústeres no tiene éxito, se repite hasta encontrar una lista de clústeres inicial valida.

4.5 RECOCIDO SIMULADO MULTI ARRANQUE

La metaheurística de recocido simulado multi arranque (MSSA) se utilizó como estrategia de búsqueda en el espacio de soluciones que se generan en el enfoque

clusterwise. El funcionamiento de esta metaheurística se presenta en el **Pseudocódigo 3** y posteriormente se explica en detalle.

```
Entrada:    Solución inicial listClusters,
            Dataset data,
            Tiempo total time,
            Numero de arranques (inicios) nBoot,
            Lista de variables a ignorar listIgnore
Salida:     Mejor solución con los modelos de regresión para los k clústeres
1   s = generarSolucionModelos (listClusters, data, listIgnore)
2   sBest = s
3   timeBoot = time / nBoot
4   para boot = 0 mientras boot < nBoot hacer
5       t = 1; tMin = 0,05; tMax = 1
6       e= exp ((log (tMin) – log (tMax)) / timeBoot)
7       timeInitial = tiempoActualSistema ()
8       timeRelease = 0
9       mientras timeRelease < timeBoot hacer
10          listClusters2 = obtenerVecino (listClusters)
11          s1 = generarSolucionModelos (listClusters2, data, listIgnore)
12          si esValidaSolucion (s1) entonces
13              si esMejor (s, s1) o aleatorio (0,1) < exp ((f(s) - f(s1)) / t) entonces
14                  listCluster = listCluster2
15                  s = s1
16                  si esMejor (sBest, s) entonces
17                      sBest = s
18                  fin si
19              fin si
20          fin si
21          t = t * e
22          timeRelease = tiempoActualSistema () – timeInitial
23      fin mientras
24  fin para
25  retornar sBest
```

Pseudocódigo 3. Algoritmo MSSA

MSSA parte de un solución inicial construida en la fase de barajado inicial (método generarSolucionModelos), esta se toma como la mejor solución **sBest** (línea 2) y la solución actual **s** (línea 1). Después en un proceso iterativo por medio de múltiples reinicios (líneas 4 a 24) realiza la búsqueda de una nueva solución **s1** denominada solución vecina generada a partir de la solución actual **s**, si esta solución **s1** (línea 11) tiene mejor *fitness* que **s**, está la reemplaza o se genera un número aleatorio con un distribución uniforme entre 0 y 1, y si este es inferior al valor de la distribución de Boltzman (línea 13), la remplaza (líneas 14 y 15). Después se compara si la nueva solución **s** supera el valor del *fitness* de **sBest** (línea 16), **s** se convierte en el nuevo **sBest** (línea 17). Finalmente se actualiza ***t*** usando un factor geométrico (línea 21) y al cumplir el criterio de parada (tiempo de ejecución) se retorna la mejor solución **sBest**.

La solución inicial es una evaluación de los grupos generados por el algoritmo k-means, esta solución tiene una estructura como se presenta en la **Figura 6**, donde

se construye para cada grupo un modelo de regresión lineal múltiple y está tiene asociado un *fitness* acorde a los *K* modelos de la solución.

El criterio de parada de esta metaheurística se estableció en un tiempo determinado para todo el proceso, el cual se dividió en el número de arranques (parámetro nBoot), esto con el objetivo de lograr en el proceso de búsqueda reinicios en la temperatura y salir de óptimos locales buscando encontrar mejores soluciones. Teniendo en cuenta lo anterior si se definen 60 minutos para la ejecución de la metaheurística con 4 arranques, el tiempo por arranque será de 15 minutos, durante los cuales la temperatura ***t*** varía según la función definida en la **Ecuación 5.**

$$t = t * e^{\frac{log\ 0.05 - \log 1}{timeBoot}} \qquad 5$$

La **Ecuación 5** genera un decrecimiento geométrico en la temperatura y esta se comporta conforme la **Figura 10**. Se debe resaltar que la temperatura desempeña un papel importante en la metaheurística de Recocido Simulado como estrategia para salir de óptimos locales, debido a que cuando la temperatura toma valores cercanos a la temperatura máxima (**tMax)** hay más probabilidad de aceptar soluciones que no son buenas, es decir realizar exploración, pero cuando decrece esta probabilidad disminuye y tiende al valor mínimo (**tMin**) realizando una explotación del vecindario ya que en general sólo se aceptan soluciones mejores a la solución actual.

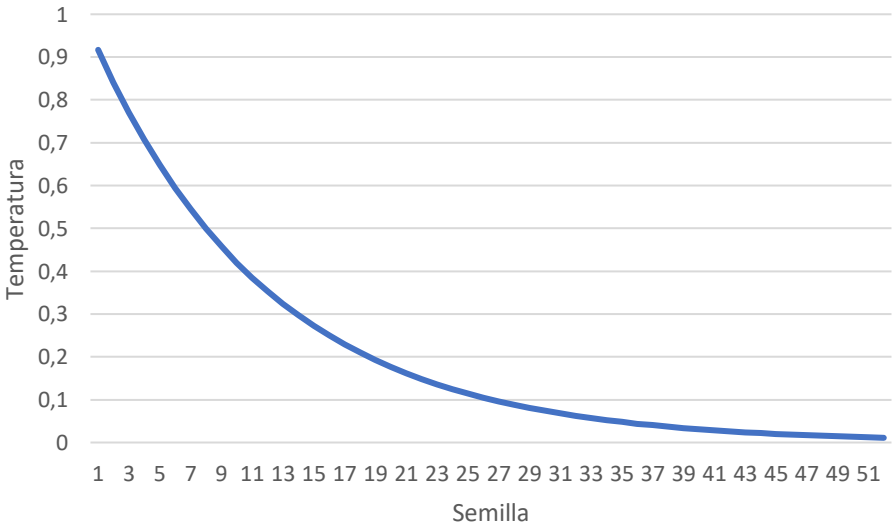


Figura 10. Comportamiento de la temperatura

El proceso de obtener un vecino (**obtenerVecino**) crea una nueva lista de clústeres (**listClusters2**) generada a partir de algunos cambios hechos sobre una copia de la lista de clústeres (**listClusters**) de la solución actual **s**. Los cambios realizados para la generación de un nuevo clúster se describen a continuación:

- Paso 1: Se elige de cada clúster de forma aleatoria el 5% de las observaciones y se ingresan en una lista de listas para tenerlas en cuenta al finalizar este paso. Supongamos que las 799 observaciones están distribuidas en 5 clúster y cada clúster tiene asignadas las siguientes cantidades de observaciones [180, 159, 175, 160, 125], teniendo en cuenta lo anterior se define una lista con 5 listas temporales con el siguiente número de observaciones [9, 8, 9, 8, 7] las cuales se eligen de forma aleatoria de los clústeres.
- Paso 2: Se recorre cada lista temporal de la lista de listas y se va asignando cada observación de forma aleatoria a un clúster que sea diferente del cual se extrajo.

La nueva lista de clústeres se usa para la construcción de una nueva solución del enfoque clusterwise, está nueva solución para ser aceptada debe cumplir el criterio de calidad conforme se explicó previamente en la **sección 4.1.1.2,** esto se realiza para que las nuevas soluciones **S1** generadas a partir de **S** sean soluciones validas (factibles) aunque no sean mejores.

La comparación de la nueva solución **S1** con la solución actual **S**, para determinar si es mejor, se hace teniendo en cuenta los modelos de cada solución, esto quiere decir que aparte del *fitness* de la solución, se compara el menor R^2 *ajustado* de cada solución y con esto se asegura que los modelos de cada solución tengan menor varianza y no se tome la decisión solo por la media del R^2 *ajustado* que se ve afectado por los valores extremos, a continuación se explica este proceso de comparación con un ejemplo.

Suponga que se tienen dos soluciones con K = 2, como se presenta en la **Figura 11**, la solución actual **s** tiene un *fitness* igual a la solución nueva **s1** con la que se está comparando, teniendo en cuenta este caso el *fitness* de una solución basado en el valor medio de los R^2 *ajustados* no es un criterio de decisión suficiente.

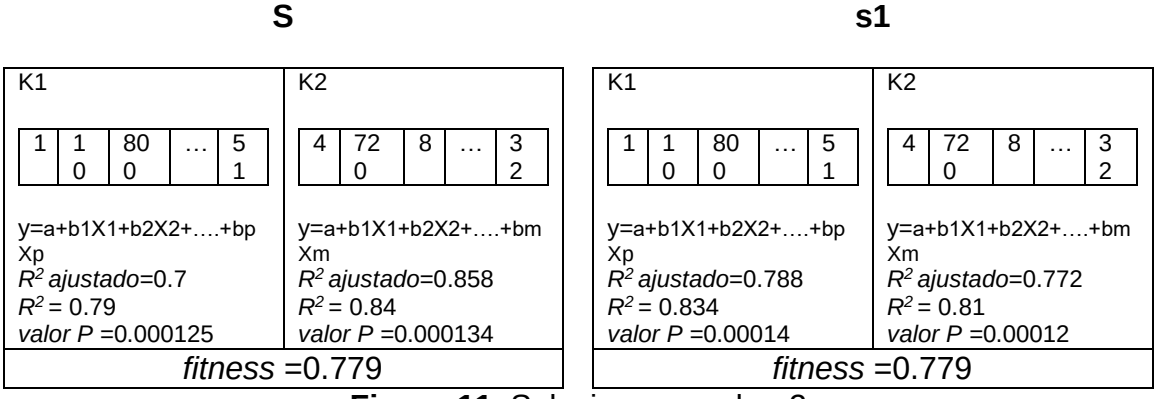


Figura 11. Soluciones con k = 2

Si una nueva solución (**s1**) es muy similar a la actual (**s**), esto es si su *fitness* está alrededor de +/- 0.1 (o un parámetro que se puede establecer previamente) el valor del *fitness* de la actual **s**, se considera que son prácticamente iguales y se agrega un nuevo criterio de comparación seleccionado como mejor solución aquella que

tenga el valor R^2 *ajustado* más alto del peor clúster de cada solución (el mayor del menor).

Teniendo en cuenta el rango antes mencionado para la solución **s** de la **Figura 11** que tiene un *fitness* de 0.799, el rango estaría comprendido desde 0.679 a 0.879, como la nueva solución **s1** está dentro del rango establecido para poder aplicar el segundo criterio de comparación se toman los valores de los modelos con menor R^2 *ajustado* de cada solución. El menor R^2 *ajustado* de la solución **s** corresponde al clúster k1 con 0.7, y el de la solución **s1** corresponde al clúster k2 con 0.772, por tanto, al comparar los R^2 *ajustados* se define que **s1** es mejor solución que **s** y de esta forma se asegura que **s1** tienen modelos con ajustes más homogéneos alrededor de la media (fitness).

Cuando la nueva solución **s1** presenta un valor de *fitness* que no está comprendido en el rango mencionado se efectúa la evaluación solo por el valor del *fitness* esto quiere decir que la solución **s1** será mejor que **s**, según el primer criterio, si su valor de *fitness* es realmente superior al valor de la solución actual por ejemplo 0.9 comparado con 0.779.

Finalmente, si una solución vecina **s1** reemplaza a la solución actual **s**, esta solución se compara con la mejor solución **sBest** encontrada durante el proceso iterativo y si esta nueva solución actual es mejor pasa a ser el nuevo **sBest**. La solución **sBest** es la que se retorna al alcanzar el criterio de parada**.**

4.6 PROCEDIMIENTO DE BÚSQUEDA CODICIOSA ALEATORIZADA Y ADAPTATIVA

La metaheurística GRASP se usó en forma similar que MSSA con la finalidad de apoyar el enfoque clusterwise a encontrar mejores soluciones. El funcionamiento de esta metaheurística se presenta en el **Pseudocódigo 4.**

Como se describió previamente en la **sección 2.1.5.2**, GRASP está compuesto por dos fases, las cuales fueron adaptadas para resolver el problema de clusterwise de la siguiente forma:

Fase 1: La fase de construcción de una solución utiliza una función codiciosa e inteligente (línea 6) encargada de crear una solución inicial (**listCluster**) generada por el algoritmo k-means y seleccionar de cada clúster el 20% de las observaciones que estuviesen más alejadas del centroide, y de esta selección elegir el 20% de forma aleatoria para realizar un intercambio a otros clústeres.

En la **Figura 12** se presenta un ejemplo de un clúster con 30 observaciones las cuales están de color verde y azul, además el círculo de color amarillo es el centroide del clúster. Teniendo en cuenta esta figura, las seis observaciones de color azul corresponden al 20% de las observaciones que están más alejadas del centroide, de estas se elige de forma aleatoria al 20% de ellas, en este caso una (1) y se intercambia a otro clúster.

```
Entrada:    Solución inicial listCluster,
            Dataset data,
            Tiempo total time,
            Fracción de tiempo (inicios) ft,
            Máximo de iteraciones mi,
            Lista de variables a ignorar listIgnore
Salida:     Mejor solución con los modelos de regresión para los k clústeres
1   sBest = vacío
2   timeMax = time / ft
3   timeInitial = tiempoActualSistema ()
4   timeRelease = 0
5   mientras timeRelease < time hacer
6       listCluster2 = contruccionCodiciosaAleatoria (listCluster, data, listIgnore)
7       si sBest == vacio entonces
8           sBest = generarSolucionModelos (listCluster2, data, listIgnore)
9       fin si
10      s = busquedaLocal (listCluster2, data, listIgnore, timeMax, mi)
11      si esMejor (sBest, s) entonces
12          sBest = s
13      fin si
14      timeRelease = tiempoActualSistema() – timeInitial
15  fin mientras
16  retornar sBest
```

Pseudocódigo 4. Algoritmo GRASP

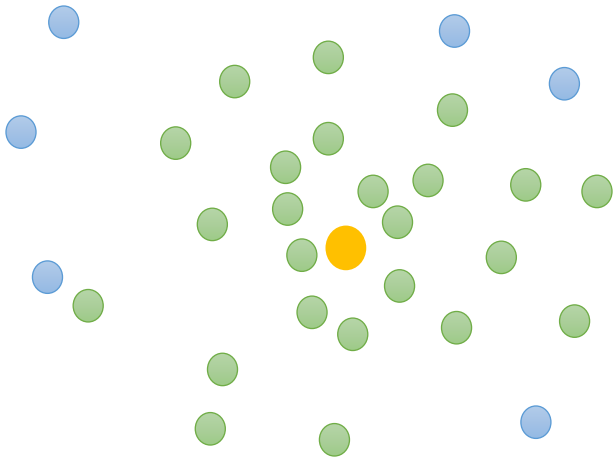


Figura 12. Ejemplo de la distribución de un clúster

Durante la selección de las observaciones que están más alejadas de los centroides, estas se extraen de cada clúster y se ingresan en una lista temporal, después el centroide del clúster afectado debe ser recalculado.

El proceso de intercambio de observaciones se ejecuta al finalizar la selección de las observaciones de todos los clústeres y la elección del clúster de destino de cada observación se hace usando la distancia euclidiana mixta para determinar el clúster más similar excluyendo el clúster de origen. Al igual que al retirar una observación de un clúster, al asignarla a otro también se realiza el recalculo del centroide del clúster afectado.

Al terminar el proceso se obtiene una nueva lista de clúster que se validó usando el criterio expuesto en la **sección 4.1.1.2** y de esta manera se asegura que la fase de construcción obtenga una solución viable para ser optimizada en la fase 2.

Fase 2: La fase de optimización local toma como punto de partida la lista de clúster construida en la fase 1 con el objetivo de encontrar una mejor solución en un tiempo determinado que corresponde a una fracción del tiempo total de ejecución de la metaheurística GRASP. El funcionamiento de la búsqueda local se presenta en **Pseudocódigo 5** y luego se explican los aspectos más relevantes de ésta.

```
Entrada:    Solución inicial listClusters,
            Dataset data,
            Lista de variables a ignorar listIgnore,
            Tiempo total time,
            Máximo de iteraciones mi
Salida:     Solución con k clusteres y sus correspondientes modelos de regresión
1   sBest = generarSolucionModelos (listClusters, data, listIgnore)
2   s = vacío
3   timeInitial = tiempoActualSistema ()
4   timeRelease = 0
6   Contador = 0
7   mientras timeRelease < time && stopCriterion != falso hacer
8       listClusters2 = obtenerVecino (listClusters)
9       s = generarSolucionModelos (listClusters2, data, listIgnore)
10      si esValidaSolucion (s) entonces
11          si esMejor (sBest, s) entonces
12              listCluster = listCluster2
13              sBest = s
14              contador = 0
15          si no
16              contador = contador + 1
17          fin si
18      si no
19          contador = contador + 1
20      fin si
21      si contador >= mi entonces
22          romper mientras
23      fin si
24      timeRelease = tiempoActualSistema () – timeInitial
25  fin mientras
26  retornar sBest
```

Pseudocódigo 5. Búsqueda local aplicada en GRASP

La búsqueda local toma inicialmente a **sBest** como la solución generada a partir de la lista de clúster construida en la fase 1, seguido se comienza un proceso iterativo de búsqueda de una mejor solución dentro del vecindario, teniendo como criterio de parada un tiempo máximo (línea 7) o de convergencia donde el algoritmo después de un numero de iteraciones definidas no encuentre soluciones con un *fitness* superior a la mejor solución.

El proceso de obtener una solución vecina a partir de una solución actual es el mismo usado en la metaheurística MSSA (línea 9) al igual que el uso de las

diferentes medidas de calidad para la validación y evaluación de las soluciones construidas.

Finalmente, después de ejecutar las dos fases antes descritas se retorna la mejor solución (**sBest**) al alcanzar el tiempo máximo de ejecución.

4.7 APORTES

En la **Tabla 7** se presenta en resumen los aportes y modificaciones hechas en los algoritmos y metaheurísticas usadas en este capítulo.

Tabla 7. Aportes adaptación enfoque clusterwise

Tema	**Aporte**
Distancia euclidiana	Se adaptó para hacer cálculo de distancia entre observaciones con datos mixtos evaluando de diferente forma las variables continuas de las nominales y binarias, estas últimas la distancia se calculó a partir de la comparación entre cadenas de la siguiente manera: • si eran iguales la distancia es 0. • si eran diferentes la distancia es 1. También para excluir del cálculo de distancias variables que tuvieran identificadores y variables de clase.
K-Means	Se hizo una adaptación en el centroide con el objetivo de optimizar el proceso de intercambio de observaciones entre clústeres y disminuir el número de operaciones, se almaceno en su interior una estructura de control que consta de lo siguiente: • Para cada variable continua tiene una sumatoria de los valores de las observaciones que pertenecen al clúster y su promedio. • Para cada variable nominal y binaria el conteo de ocurrencias de los valores que tiene por cada observación.
Regresión lineal múltiple	Se utilizó la librería de weka para hacer cálculo de las regresiones lineales para cada clúster de una solución. Se realizó un wrapper para paso de datos entre las estructuras propias y las estructuras requeridas por weka. Se realizaron algunas modificaciones a algunas clases de las librearías de weka para hacer cálculos adicionales necesarios como cálculo del valor P, tener acceso a los

	valores de la regresión, variables seleccionadas en la regresión.
MSSA	Se adaptó para el funcionamiento por arranques, es decir que el criterio de búsqueda era ejecución en un tiempo determinado, por tanto se dividió este tiempo en n arranques en donde la función de temperatura permitía reiniciarse y no caer en óptimos locales. En el proceso de búsqueda de nuevas soluciones o soluciones en el vecindario el objetivo fue intercambiar observaciones entre clústeres. La función objetivo consistió en la evaluación de regresiones lineales múltiples por cada clúster y que finalmente a partir de los R_2 ajustados de cada clúster hiciera un promedio que definiría la calidad de una solución y esta se buscaba maximizar.
GRASP	El construcción de la lista restringida de candidatos, se construyó a partir de un proceso de clustering con k-means y de este proceso a cada clúster se eligió un porcentaje de observaciones las cuales estuvieran más alejadas del centroide de las cuales se elegiría de forma aleatoria algunas para un proceso de intercambio a otros clústeres y de esta manera construir la lista de soluciones. En la búsqueda local se adaptó para tener como función objetivo el uso de regresiones lineales múltiples y maximizar el R_2 ajustado promedio, adicionalmente otro criterio también usado para evitar estancamientos en la búsqueda fue identificar si tras n iteraciones no se mejoraba la solución se terminaba la búsqueda local.

4.8 EXPERIMENTACIÓN

Este ítem se encuentra organizado en varias secciones, explicando en cada una de ellas, diferentes partes de los algoritmos y de los datos que se debieron definir. Para terminar, se presentan los resultados obtenidos en el experimento final con la versión completa y estable de los datos y los algoritmos.

En cuanto a los algoritmos se requirió definir la medida de distancia que se debía usar, definir el criterio de calidad (fitness) de una solución en el enfoque clusterwise, realizar el afinamiento de los algoritmos (k-means, MSSA y GRASP) y el escenario de los experimentos para poder comparar en forma justa los resultados. Durante el

desarrollo del proyecto se realizaron varias mejoras a la vista minable que también se explican en este apartado del documento.

La ejecución de los algoritmos se realizó en dos equipos, cuyas características son las siguientes:

- Equipo 1: procesador AMD PHENON X4 a 3.0 GHz, 16 GB RAM, HDD 7200 RPM y sistema operativo Windows 10.
- Equipo 2: procesador Intel Core i7 a 2.4, 8GB RAM, SSD Kingston uv400 y sistema operativo Windows 10.

Durante el proceso de experimentación se ejecutaron 90 pruebas que comenzaron desde el mes de junio del año 2018 y culminaron en el mes de noviembre, estas se realizaron de forma transversal a la construcción de los diferentes algoritmos junto con el afinamiento de los respectivos parámetros.

4.8.1 Definición de la distancia

Primero se definió cual de dos tipos de distancias para variables mixtas era la más apropiada para el problema concreto de la tesis, la distancia de Gower o la distancia euclidiana mixta. Esta última fue elegida por tener mejor desempeño en la construcción de la solución inicial en el algoritmo k-means. La elección se tomó teniendo en cuenta que el error cuadrático medio (SSE) asociado a los modelos construidos para cada agrupación fue inferior comparado con la distancia de Gower, además de generar soluciones de agrupamiento factibles en menor tiempo.

4.8.2 Definición del criterio de calidad (fitness) de una solución

Teniendo en cuenta el estado del arte (clusterwise aplicado a la gestión de pavimentos presentado en la sección 2.2.2) se evaluó el criterio bayesiano de información (Bayesian Information Criterion, BIC) formalmente representado por la **Ecuación 6**

$$BIC = n + n * \ln(2\pi) + n * \ln\frac{SSE}{n} + (\delta + K - 1) * \ln n \qquad 6$$

Donde ***n*** es el número de observaciones del modelo, ***SSE*** es la suma total de errores cuadráticos, δ es el número total de variables explicativas del modelo, K número de clústeres.

Partiendo de la **Ecuación 6**, se realizó la adaptación del BIC para medir la calidad del macro modelo de la solución del enfoque, teniendo en cuenta que una solución está compuesta por *k* modelos y cada modelo tiene asociado una suma de cuadrados *SSE*, un número de variables explicativas δ y *n* observaciones.

Se ejecutaron 30 ejecuciones de 2 hasta 5 clústeres, los resultados obtenidos se presentan en la **Tabla 8**, según la interpretación del BIC la mejor solución corresponde al menor valor, esto quiere decir que la mejor solución corresponde a la ejecución con 2 clústeres y semilla 35.

Tabla 8. Resultados ejecución BIC solución inicial

Semilla	BIC K2	BIC K3	BIC K4	BIC K5
1	-12.405	-6.092	0.695	7.550
3	-12.539	-6.024	0.840	7.464
5	-12.434	-6.088	0.575	7.955
7	-12.393	-5.988	0.658	7.451
9	-12.317	-6.046	0.631	7.331
11	-12.428	-5.990	0.690	7.465
13	-12.482	-5.992	0.757	7.104
15	-12.437	-5.774	0.503	7.468
17	-12.453	-6.172	0.611	7.777
19	-12.367	-6.016	0.730	7.824
21	-12.426	-5.954	0.678	7.523
23	-12.349	-6.136	0.597	7.511
25	-12.195	-5.971	0.544	7.466
27	-12.318	-6.027	0.725	7.547
29	-12.554	-5.963	0.551	7.631
31	-12.426	-5.924	0.568	7.505
33	-12.406	-6.149	0.861	7.699
35	**-12.621**	-5.925	0.856	7.614
37	-12.501	-5.908	0.627	7.423
39	-12.378	-5.913	0.652	7.535
41	-12.329	-6.051	0.657	7.628
43	-12.397	-5.949	0.588	7.282
45	-12.546	-5.771	0.811	7.383
47	-12.585	-6.096	0.942	7.909
49	-12.178	-5.908	0.531	7.205
51	-12.329	-5.793	0.749	7.294
53	-12.328	-5.961	0.819	7.751
55	-12.519	-5.912	0.758	7.634
57	-12.592	-5.590	0.585	7.509
59	-12.445	-5.716	0.632	7.856

Contrastando los valores del BIC con otros valores de calidad asociados a los modelos obtenidos, como *SSE, R^2, R^2 ajustado*, se identificó que el valor del BIC en las mejores soluciones no correspondía con los mejores modelos según los otros valores de calidad, lo cual llevó a analizar el desempeño de las otras medidas calidad e inclusive analizar otra fórmula de BIC [42] reportada en la literatura, y realizar la adaptación de la función log-likelihood acorde a cada modelo de una solución como se presenta en la ecuación 7, pero sin obtener buenos resultados.

$$BIC = -2 * \ln \sum_{i=1}^{k} \frac{SSE_i}{no_i} * \frac{vs_i}{v_i} + K * \ln n \qquad 7$$

Donde ***n*** es el número de observaciones del modelo, ***SSE$_i$*** es la suma total de errores cuadráticos del clúster *k*, ***no$_i$*** número de observaciones del clúster *i*, ***vs$_i$*** numero de variables seleccionadas en el clúster *i*, ***v$_i$*** número total de variables del clúster *i*, *K* número de clústeres.

Los análisis realizados a los resultados de estas pruebas con base en los mejores modelos tomando como criterio de calidad el promedio del *R^2 ajustado* desde k=2 hasta k=5 se pueden apreciar en la **Figura 13,** donde los valores de *R^2 ajustado* de cada modelo están en color azul y el valor promedio de estos está en color amarillo; la mejor solución usando este criterio corresponde a una ejecución de k=4 y no a k=2 como lo reportó el BIC.

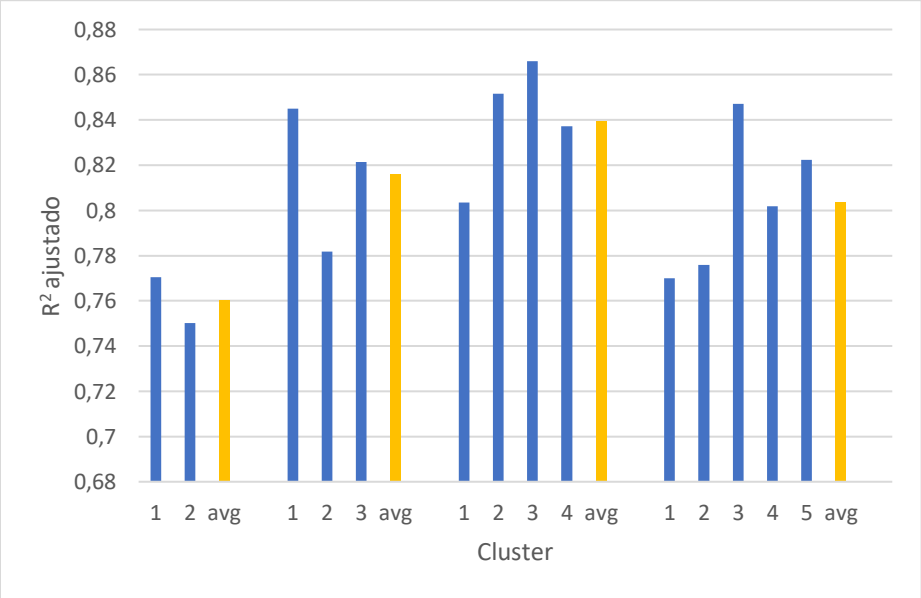


Figura 13. Grafica de R^2 ajustado de los mejores modelos con k = 2 hasta 5

Entrando en más detalles se analizó otra característica de los modelos como los grados de libertad, llegando a la conclusión que el BIC planteado en la **Ecuación 6** al usarse en macro modelos castiga mucho una solución cuando está compuesta por más de 3 clústeres y no tiene en cuenta el error y los grados de libertad, por lo tanto, no se considera un criterio que aporte en la elección de la mejor solución del enfoque clusterwise para los datos tratados en el proyecto.

Partiendo de lo anterior se tomó la decisión de usar en su lugar las medidas de calidad propias de los modelos, específicamente el promedio de *R^2 ajustado* teniendo en cuenta que es una medida que incorpora el número de parámetros del modelo, el número de observaciones y el error asociado al modelo. Esta medida de calidad fue más confiable y apropiada para elegir los macro modelos del enfoque

clusterwise en los datos del proyecto, sabiendo que se van a comparar con otras soluciones de diferente número de clústeres.

4.8.3 Mejoras en la vista minable (dataset)

Los datos usados para la construcción de los modelos fueron el dataset inicial y 3 versiones nuevas, generadas a partir de la aplicación de diferentes tratamientos para la reducción de dimensionalidad descritos previamente en la **sección 3.3.3.**

Durante la ejecución de las primeras evaluaciones del proceso de optimización usando las metaheurísticas MSSA y GRASP se logró identificar que los mejores modelos fueron construidos usando el dataset inicial, este comportamiento se debe a que algunas variables que no están presentes en los 3 dataset derivados aportan información a los modelos de algunos clústeres y esto hace que el macro modelo tenga mejor calidad según la medida definida R^2 *ajustado* promedio.

Después de culminada la etapa de construcción de modelos por clusterwise y en la fase optimización de los rendimientos (presentada en el capítulo siguiente) se presentaron algunos casos especiales donde un experto del CIAT sugirió realizar un cambio en el dataset inicial con la finalidad de disminuir complejidad en los modelos de regresión lineal, lo que dio origen a una nueva iteración para la construcción de los nuevos modelos con la nueva versión del dataset inicial.

4.8.4 Definición de un escenario para la experimentación

En un comienzo la ejecución de las diferentes pruebas se realizó de manera secuencial con el dataset inicial con las dos metaheurísticas propuestas (MSSA y GRASP), construyendo macro modelos con *k* clústeres (desde 2 hasta 5). Esta construcción usó como criterio de parada un número de iteraciones máximo, después de varias pruebas se optó por cambiar el criterio de parada a un tiempo total o máximo de ejecución.

La anterior decisión se tomó basada en el resultado de los experimentos debido a que al usar un número máximo de iteraciones para k=2, k=3 las ejecuciones terminaban mucho más rápido (menos tiempo) que con k=4 y k=5 ya que la evaluación de la función objetivo (crear los modelos y sacar las estadísticas apropiadas para el cálculo del fitness) es más costosa cuando el valor de *k* crece.

Al realizar ejecuciones secuenciales y elegir un tiempo total para cada ejecución de las metaheurísticas de 12 horas por *k* (desde 2 hasta 5), estas tendrían una duración total de 48 horas equivalentes a 2 días para cada solución propuesta, teniendo en cuenta que se necesitaba ejecutar esta prueba al menos 30 veces con diferentes semillas. Este proceso se demoraría 60 días por metaheurística para un total de 120 días.

Partiendo de los cálculos anteriores del tiempo de ejecución, se paralelizó la ejecución de las metaheurísticas usando diferentes hilos y se logró aprovechar al máximo la capacidad computacional de los equipos, por tanto, el tiempo de pruebas pasó de los 60 días de ejecución estimados por cada metaheurística a un total de

dos días en los cuales se ejecutan las 30 ejecuciones con diferentes semillas en 30 hilos por separado.

La optimización en los tiempos de ejecución facilitó el proceso de afinamiento de parámetros usados en las metaheurísticas, además sirvió para identificar que después de cierto tiempo en la mayoría de las ejecuciones para cada *k* (desde 2 hasta 5) la mejora en las soluciones no era considerable, encontrando un balance entre el tiempo de ejecución por número *k* de grupos entre 3 y 6 horas.

El control adecuado de la generación de números aleatorios a partir de una semilla permitió asegurar la repetibilidad de los experimentos bajo las mismas condiciones, esto quiere decir, que los resultados son iguales si se usan las semillas iguales, se hace en una maquina con las mismas condiciones y se usa el mismo tiempo de ejecución. Las semillas usadas fueron los números impares del 1 hasta el 59.

4.8.5 Afinamiento del algoritmo K-means

Los parámetros que se afinaron en k-means fueron el tipo de distancia y el número de iteraciones. Teniendo en cuenta que, para la construcción de un modelo de regresión lineal, se requiere que el número de observaciones sea mayor al número de variables, cuando se usaban valores de *k* superiores a 5 no se cumplía con dicha restricción, por eso, los algoritmos se ejecutaron con *k* desde 2 clústeres hasta 5.

El número de iteraciones máximas *n* usadas por k-means para converger fue de 500, teniendo en cuenta que para cuando el valor de *k* supera el valor de 4, es más complejo encontrar combinaciones que sean factibles y se pueden generar los modelos para todos los clústeres.

La distancia utilizada fue la euclidiana gracias a su desempeño, además se ignoraron las variables de clase e identificador de cada observación en el cálculo de distancias. A continuación, se resumen los parámetros usados en k-means.

Numero de clústeres	K = 2, … 5
Numero de iteraciones	N = 500
Tipo de distancia	tDis = 1 (euclidiana=1, gower=0)
Lista de variables a ignorar	listIgno = [RDT_AJUSTADO, ID_LOTE]
Ignorar variables	Ignorar = true

4.8.6 Afinamiento del algoritmo MSSA

La solución inicial de la metaheurística MSSA se construye con el algoritmo k-means, que se construye con *k* clúster que pueden ser 2, 3, 4 o 5. Esta metaheurística tiene un comportamiento particular y es que al ser multi arranque los resultados dependen del tiempo de ejecución dado y el número de arranques definidos, para este caso se determinaron 5 arranques, en los cuales la temperatura varía en cada arranque en un rango previamente definido desde un valor máximo a uno mínimo con un decrecimiento geométrico como se comentó previamente, también se utilizó un decrecimiento lineal pero no fue tan bueno.

El tiempo utilizado para la ejecución fue de 3 horas para pruebas desde k = 2 hasta k = 5. Teniendo en cuenta los resultados obtenidos con k = 5, finalmente se le asignó un tiempo mayor, de 6 horas y las pruebas finales se ejecutaron sobre el equipo 1.

En la búsqueda de soluciones vecinas se tomó la decisión de realizar intercambios del 5% de observaciones de cada clúster con el objetivo de que se realice una explotación sobre el vecindario de la solución actual. A continuación, se resumen los parámetros que se usaron en la experimentación. Es preciso comentar que se necesita realizar un proceso más formal de afinamiento de parámetros del algoritmo para obtener aún mejores resultados.

Tiempo total (milisegundos)	time=10800000 … time=21600000
Numero de arranques	nBoot=5
Lista de variables a ignorar	listIgno=[RDT_AJUSTADO, ID_LOTE]
Porcentaje de cambios	pCambios=0.05

4.8.7 Afinamiento del algoritmo GRASP

El proceso de construcción codiciosa (fase 1 de GRASP) toma como punto de partida la solución construida por el algoritmo k-means e identifica en cada clúster el 20% de las observaciones más alejadas del centroide como candidatas para intercambio a otros clústeres, de las cuales se elige de forma aleatoria el 20% para su intercambio, estos porcentajes se definieron con el objetivo de realizar una construcción codiciosa rápida donde se logra una explotación de la solución actual y se consigue una nueva distribución de observaciones válida en los clústeres. Con estos porcentajes se obtenían soluciones diferentes y válidas, dando un nivel de exploración adecuado sin incrementar el tiempo requerido para esta fase.

La metaheurística en la búsqueda local (fase 2) utiliza el mismo criterio de MSSA para encontrar nuevas soluciones, intercambiando el 5% de las observaciones de cada clúster con la finalidad de realizar una explotación en el espacio de búsqueda de la solución inicial construida, además de utilizar doble criterio de parada definido en un tiempo máximo de ejecución o un criterio de convergencia que no supere un máximo de 20 iteraciones sin encontrar mejoras. La fracción de tiempo que se utilizó para la ejecución de la búsqueda local fue de 10, esto quiere decir que el tiempo máximo de la búsqueda local fue de la décima parte del tiempo total, si es de 60 minutos solo se ejecutó por 6 minutos.

El tiempo utilizado para la ejecución del algoritmo fue de 3 horas para pruebas desde k=2 hasta k=5; teniendo en cuenta que los resultados al igual que MSSA fueron mejores para k=5, la metaheurística se ejecutó por un tiempo total de 6 horas y las pruebas finales se ejecutaron sobre el equipo 1. A continuación, se resumen los parámetros del algoritmo. Para este algoritmo también se necesita realizar un proceso más formal de afinamiento de parámetros del algoritmo.

Tiempo total (milisegundos)	time = 10800000 … time = 21600000
Fracción de tiempo	ft = 10
Lista de variables a ignorar	listIgno = [RDT_AJUSTADO, ID_LOTE]
Porcentaje de cambios	pCambios = 0.05
Máximo de iteraciones	mi=20

4.9 COMPARACIÓN DE RESULTADOS

Cada algoritmo (MSSA, GRASP) se ejecutó 30 veces (valor apropiado para tener una media y desviación que tiende a tener una distribución normal según el teorema del límite central), con 2, 3, 4 y 5 grupos. Una vez ejecutados, se calcularon los R^2 *ajustados promedios* por número de clústeres, estos se pueden observar en la **Tabla 9** (mejores resultados por cada algoritmo en negrita, 3 horas de ejecución para cada clúster). Las dos adaptaciones de MSSA y GRASP hallaron un mayor R^2 *ajustado promedio* cuando *K* es igual a 5.

Tabla 9. R^2 ajustado promedio para 2, 3, 4 y 5 clústeres

K	MSSA R^2 Ajustado promedio	GRASP R^2 Ajustado promedio
2	0.768526008	0.779104195
3	0.800398698	0.814187886
4	0.832382658	0.844995661
5	**0.859444174**	**0.862710015**

4.9.1 Comparación por solución propuesta con diferentes valores de *k*

Para poder identificar si alguna de las dos soluciones propuestas presentaba un mejor desempeño, se realizó la prueba estadística no paramétrica de Wilcoxon para dos muestras apareadas (**ANEXO 9**), esto con la finalidad de conocer si existía una diferencia estadísticamente significativa entre las medias de los resultados obtenidos.

La prueba de Wilcoxon fue diseñada para comparar dos muestras de las cuales no se puede definir a priori la distribución de los datos debido a la cantidad mínima de eventos capturados, la prueba busca conocer si las diferencias entre las dos muestras son debido al azar o no.

El resultado de la prueba de Wilcoxon sobre los valores de la **Tabla 9** arrojó que no existía una diferencia estadísticamente significativa entre las medias de los R^2 *ajustados* obtenidos en las regresiones de las soluciones propuestas, por lo que no hay evidencia de que la adaptación desarrollada con GRASP o MSSA presenta mejor desempeño.

Debido a lo anterior y observando que ambas soluciones propuestas presentaban un mayor R^2 *ajustado promedio* al trabajar con 5 clústeres, se tomó la decisión de identificar si dentro de estas ejecuciones podía existir una diferencia estadísticamente significativa.

4.9.2 Comparación por solución propuesta con 5 grupos

La prueba de Wilcoxon tiene como su alternativa paramétrica a la prueba *t* para muestras apareadas. Como se cuenta con 30 valores, resultado de 30 nuevas ejecuciones para *K* = 5 con el doble del tiempo, 6 horas, se utilizaron estas dos pruebas para determinar la diferencia entre las medias de las muestras. Los R^2 *ajustados promedios* se pueden observar en la **Tabla 10** (mejores modelos encontrados por cada algoritmo en negrita, con 6 horas de ejecución por cada resultado).

Tabla 10. R^2 ajustado promedio para 5 clústeres

MSSA	**GRASP**
0.85896178	0.86699366
0.86735142	0.86789029
0.86802873	0.86541297
0.86708354	**0.87890047**
0.85282285	0.86485148
0.87399456	0.8679664
0.86044491	0.86102361
0.86408723	0.85172357
0.87643251	0.85084143
0.86577553	0.8656843
0.8527021	0.86142286
0.85663142	0.86269508
0.85885181	0.85716346
0.861856	0.8681765
0.85438244	0.85482283
0.86705984	0.87667793
0.87455223	0.87750622
0.86047963	0.86417865
0.86018813	0.85324996
0.85580508	0.86175778
0.87161131	0.87622102
0.85089143	0.86522161
0.86137075	0.85837452
0.85790498	0.86711009
0.87218249	0.86712982
0.86844306	0.87383793
0.86674322	0.86750137
0.86635248	0.87572715
0.85873027	0.86640279
0.87883411	0.87078299
PROMEDIO: 0.86368519	**PROMEDIO: 0.86557496**
DESV.ESTANDAR: 0.00696649	DESV.ESTANDAR: 0.00760283

Los resultados de la prueba de Wilcoxon se pueden observar en el **ANEXO 9** y los de la prueba *t* para muestras apareadas en el **ANEXO 10**. El resultado de ambas pruebas arrojó que no existía una diferencia estadísticamente significativa entre las medias de los R^2 *ajustados promedios* obtenidos por las soluciones encontradas con $K = 5$ en los dos algoritmos.

4.9.3 Selección del mejor macro modelo

Al no encontrar diferencia estadísticamente significativa entre las medidas de calidad de ambas soluciones propuestas y con la necesidad de definir el mejor macro modelo (solución con 5 modelos de regresión, uno para cada clúster), se optó por seleccionar los 10 mejores R^2 *ajustados promedios* de la **Tabla 10**. Entre estos 10 modelos no se presentó una diferencia mayor al 0.01% en la calidad y se seleccionó el macro modelo más simple, debido a que un modelo con menor número de variables es más fácil de interpretar. La **Tabla 11** muestra el detalle de los 10 mejores modelos encontrados, además muestra en negrita el seleccionado.

Tabla 11. Top 10 mejores R2 ajustados promedios (macro modelo más simple en negrita)

			Numero de variables seleccionadas por Clúster					
Solución Propuesta	**Semilla**	**R^2 Ajustado promedio**	**Clúster 1**	**Clúster 2**	**Clúster 3**	**Clúster 4**	**Clúster 5**	**Promedio**
GRASP	7	0.87890047	68	64	78	54	55	63,8
MSSA	59	0.87883411	60	79	61	35	72	61,4
GRASP	33	0.87750622	67	59	57	74	61	63,6
GRASP	31	0.87667793	49	42	62	65	65	56,6
MSSA	17	0.87643251	51	73	67	47	67	61
GRASP	**41**	**0.87622102**	**46**	**68**	**48**	**59**	**57**	**55,6**
GRASP	55	0.87572715	49	51	71	80	68	63,8
MSSA	33	0.87455223	55	55	66	59	73	61,6
MSSA	11	0.87399456	55	65	74	49	67	62
GRASP	51	0.87383793	54	70	66	72	47	61,8

El macro modelo con menor promedio de variables seleccionadas fue creado por la solución propuesta con GRASP y la semilla 41, este macro modelo fue enviado para su revisión junto con los clústeres y el respectivo análisis descriptivo de cada uno de estos a Hugo Andrés Dorado Betancourt (director del proyecto) y Sylvain Jean Delerce, ambos investigadores del área de investigación en Análisis de Decisiones y Políticas (DAPA) del CIAT.

4.10 CONCLUSIONES

Al revisar los clústeres obtenidos en el marco modelo se observan características de suelo, clima y prácticas de manejo que los hacen diferentes. A continuación, se describe por cada clúster las características que lo diferencian de los demás:

Clúster 1 (Tercer mejor rendimiento promedio "4768 kg/ha")
Características de suelo: • Estructura granular • Recubrimiento vegetal espaciado • Drenaje interno bueno • Drenaje externo lento

• El suelo en su mayoría presenta una textura Far (53.82%) y dureza Friable (45.73%)
Características de clima: • Presenta las mayores frecuencias de temperatura Max de 34°C promedio en todas sus etapas (40% en las etapas vegetativa y formación; 30% en la etapa de maduración), además al interior el clúster se encuentra la mayor frecuencia de temperatura Max de 34°C en etapa de formación (78%) • Al interior del clúster se reportó la mayor energía acumulada en etapa vegetativa (16960) y menor energía acumulada en etapa de formación (19463) • Presenta mayores valores de rango diurno en grados Celsius en la etapa vegetativa y de formación (9.1, 9.5, 9.14)
Características de manejo: • Único clúster en donde se aplicó N,P y K antes de la siembra • Se adecua drenaje para el cultivo • Cerca de ¾ de los registros realizaron una siembra mecanizada • Cerca de ⅔ de los registros presentaban como cultivo anterior el algodón

Clúster 2 (Cuarto mejor rendimiento promedio “4568 kg/ha”)
Características de suelo: • Estructura granular • Recubrimiento vegetal regular • Drenaje interno lento a muy lento • Sin drenaje externo • Tiene menor profundidad efectiva (32 cm, presentando 12 cm por debajo frente al promedio de los demás clústeres) • Se observan moteados y la profundidad promedio es de 18.82cm • El suelo en su mayoría presenta una textura Far (34.75%), Ar (25.57%) y dureza Blando (61.13%)
Características de clima: • Presenta las menores frecuencias de temperatura Max de 34°C promedio en todas sus etapas (30% en la etapa vegetativa y formación; 22% en la etapa de maduración), además al igual que el clúster 1, se encuentra la mayor frecuencia de temperatura Max de 34 en la etapa formación (78%); también se destaca porque se encuentra la mayor frecuencia en etapa de maduración (91%) • La precipitación acumulada promedio en mm está entre las más bajas durante las 3 etapas (221, 247, 119), además al interior del clúster se reportó la menor precipitación acumulada en mm para la etapa vegetativa (51.4) y la etapa de formación (14.9) • La frecuencia de precipitación inferior a 10mm en promedio está entre las más bajas para las 3 etapas (16%, 14%, 10%). • Presenta el mayor promedio de humedad durante las tres etapas (83%,82%,83%) Al interior del clúster se reportó la menor energía acumulada

en la etapa vegetativa (13132) y mayor energía acumulada en la etapa de maduración (27398)
- Presenta menores valores de rango diurno en grados Celsius en las 3 etapas (8.71, 9.12, 8.77)

Características de manejo:
- Se aplicó la menor cantidad de fertilizante en general
- No se adecuá drenaje para el cultivo
- Cerca de ⅔ de los registros realizan siembra mecanizada
- Cerca de ½ de los registros presentaban como cultivo anterior el algodón

Clúster 3 (Mejor rendimiento promedio "5102 kg/ha")

Características de suelo:
- Estructura aterronada
- Recubrimiento vegetal bueno
- Drenaje interno bueno
- Sin drenaje externo
- El suelo en su mayoría presenta una textura Ar (41.85%) y dureza Firme (40.94%)

Características de clima:
- La etapa vegetativa y de maduración están entre las mayores frecuencias de temperatura Max de 34° promedio (36% y 34%), pero la etapa de maduración es de las más bajas (20%).
- La precipitación acumulada promedio en mm está entre las más altas para las 3 etapas (263, 256, 138)
- La frecuencia de precipitación menores a 10mm promedio está entre las más altas para las 3 etapas (18%,15%,13%)

Características de manejo:
- Presenta las mayores cantidades de N,K aplicadas antes de la siembra
- Único clúster en donde se reportan fertilizaciones orgánicas (entre las etapas de emergencia y de floración)
- Se adecuá drenaje para el cultivo
- Cerca de ¾ de los registros realizan siembra mecanizada
- Cerca de ⅔ de los registros presentaban como cultivo anterior el algodón

Clúster 4 (Segundo mejor rendimiento promedio "4902 kg/ha")

Características de suelos:
- Una estructura aterronada
- Recubrimiento vegetal bueno
- Se observa hojarasca
- Drenaje interno bueno
- Drenaje externo lento
- El suelo en su mayoría presenta una textura Far (54.23%) y dureza Firme (34.11%), Friable (33.4%)

Características de clima: • Al interior del clúster se reportó la mayor energía acumulada en la etapa de formación (26199) y la menor energía acumulada en etapa de maduración (2996) • Presenta mayor valor de rango diurno en grados centígrados (9.43) en la etapa de maduración.
Características de manejo: • Se adecuá drenaje para el cultivo • Cerca de ¾ de los registros realizan siembra mecanizada • Cerca de ½ de los registros presentaban como cultivo anterior el maíz

Clúster 5 (Peor rendimiento promedio "4335 kg/ha")
Características de suelo: • Estructura aterronada • Recubrimiento vegetal bueno • Drenaje interno bueno • Sin drenaje externo • El suelo en su mayoría presenta una textura FAr (29.79%), Ar (23.59%) y dureza Firme (31.75%), Duró (26.86%)
Características de clima: • Al interior del clúster se reportó la mayor frecuencia de temperatura Max de 34°C en la etapa vegetativa (82%). • La precipitación acumulada promedio en mm está entre las más bajas para las 3 etapas (212, 245, 125) • Presenta el menor promedio de humedad durante las tres etapas (80%,80%,81%) • Al interior del clúster se reportó la menor energía acumulada en etapa vegetativa (13132)
Características de manejo: • No se adecua drenaje para el cultivo • Cerca de ½ de los registros realizan siembra manual • Cerca de ½ de los registros presentaban como cultivo anterior el maíz

De la caracterización de los clústeres se logró concluir que aquellos con mejor rendimiento promedio se asocian con drenaje y fertilizaciones más altas, para el caso de los peores rendimientos promedios existe una relación con suelos pesados y mal drenados.

La distribución de las observaciones al interior de los 5 clústeres obtenidos confirma lo planteado en la Agricultura Especifica Por Sitio (AEPS), la cual indica que dos o más eventos de cultivos son similares no por su cercanía geográfica sino por las prácticas de manejo agrícola, características climáticas y de suelo que comparten. Esto se observa en la **Figura 14** y en la **Figura 15** donde se identifican por medio de colores la asignación de los eventos en los cinco (5) clústeres.

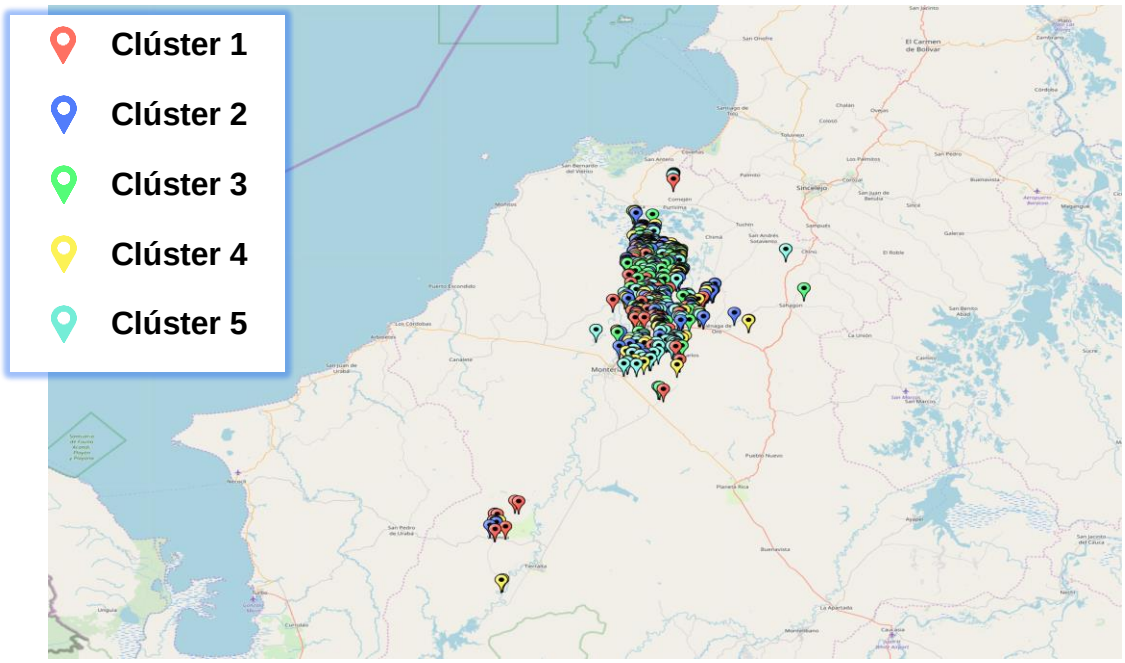


Figura 14. Distribución cultivos en clúster departamento de Córdoba – Colombia

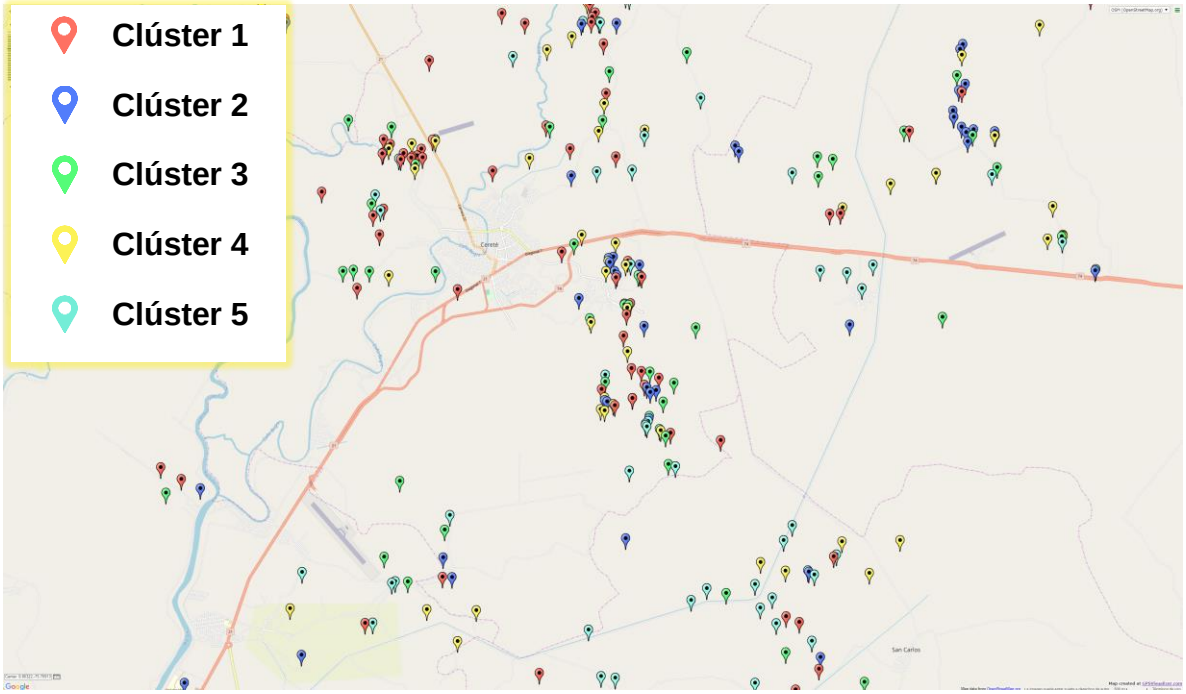


Figura 15. Distribución cultivos en clúster departamento de Córdoba – Colombia (ampliada), acercamiento a la zona de cultivos más densa

Teniendo en cuenta la distribución de los eventos en los clústeres, se realizó una comparación de los centroides de cada clúster, donde se resaltan los valores para cada variable (**Tabla 12**), con lo anterior se identificó que las mayores diferencias se presentaron en las características relacionadas con el suelo como la resistencia al rompimiento y textura del suelo. Con estas diferencias es posible realizar una

caracterización de los eventos de cultivo en términos de condiciones de suelo y diseñar una estrategia de fertilización específica por cada clúster, esto último partiendo de que la variabilidad de los nutrientes en el suelo afecta directamente el rendimiento, y si estos son aplicados de manera uniforme pierden efectividad [43].

Tabla 12. Representación de los cinco clústeres del mejor macro modelo encontrado

Variable	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
TIPO_SIEMBRA	Mecanizado	Mecanizado	Mecanizado	Mecanizado	Manual
SEM_TRATADAS	No	No	No	No	No
MATERIAL_GENETICO	Otro	Otro	Otro	Otro	Otro
CULT_ANT	Algodón	Algodón	Algodón	Maíz	Maíz
DRENAJE	Si	No	Si	Si	No
METODO_COSECHA	Manual	Manual	Mecanizado	Manual	Manual
ALMACENAMIENTO_FINCA	No	No	No	No	No
DIAS_EN_EMERGER	5.01	4.54	5.21	5.17	5.25
DIAS_EN_EMERGER_A_FLORECER	46.95	47.21	48.51	47.56	47.96
DIAS_EN_FLORECER_A_COSECHAR	80.33	81.52	78.62	82.34	81.23
POBLACION_20DIAS_AJT	65971	61645	64311	66579	61607
ALTURA_LOT	15.8	9.48	12.48	11.66	12.69
ContEnfQui_Emer_Flor	0.2	0.16	0.14	0.17	0.1
ContEnfQui_Flor_Cose	0.04	0	0.01	0.02	0.02
ContMalMec_Siem_Emer	0	0	0.01	0	0
ContMalMec_Emer_Flor	0.01	0	0	0	0.03
ContMalMec_Flor_Cose	0	0	0.01	0.01	0
ContMalQui_Antes_Siem	0.16	0.03	0.12	0.18	0.09
ContMalQui_Siem_Emer	0.71	0.51	0.36	0.47	0.33
ContMalQui_Emer_Flor	1.02	0.98	1.01	1.12	0.80
ContMalQui_Flor_Cose	0.03	0.02	0.05	0.01	0.02
ContPlaQui_Antes_Siem	0.05	0.01	0.04	0.12	0.02
ContPlaQui_Siem_Emer	0.2	0.05	0.07	0.08	0.04
ContPlaQui_Emer_Flor	1.81	1.38	1.86	1.8	1.48
ContPlaQui_Flor_Cose	0.1	0.03	0.07	0.08	0.08
TotN_Antes_Siem	0.08	0	1.32	0.63	0.34
TotN_Siem_Emer	2.15	1.14	2.78	2	1.11
TotN_Emer_Flor	92.59	81.82	87.8	90.37	103.3
TotP_Antes_Siem	0.16	0	0	0	0
TotP_Siem_Emer	2.87	0.3	1.32	1.8	0.69
TotP_Emer_Flor	1.99	2.15	1.33	2.05	1.99
TotK_Antes_Siem	0.08	0	0.18	0	0
TotK_Siem_Emer	2.65	0.47	1.75	2.18	1.10
TotK_Emer_Flor	26.27	10.89	19.3	17.82	20.49
FerOrg_Emer_Flor	0	0	0.01	0	0

Variable	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
FerQui_Antes_Siem	0.01	0	0.03	0.01	0.01
FerQui_Siem_Emer	0.22	0.05	0.15	0.16	0.09
FerQui_Emer_Flor	2.78	2.18	2.58	2.6	2.49
PENDIENTE_RASTA	2.94	1.66	2.31	2.52	2.15
TERRENO_CIRCUN_RASTA	Plano o Llano	Plano o Llano	Plano o Llano	Plano o Llano	Plano o Llano
POSICION_PERFIL_RASTA	Plano	Plano	Plano	Plano	Plano
NO_CAPAS_RASTA	2.05	2.78	2.12	2.2	2.04
PH_RASTA	5.69	5.86	5.76	5.68	5.62
PEDREG_PERFIL_ROCAS	Sin Rocas	Sin Rocas	Sin Rocas	Sin Rocas	Sin Rocas
CAP_ENDURE_RASTA	No	No	No	No	No
PROFUND_CAP_ENDURE_RASTA	-1	-1	-1	-1	-1
ESPESOR_CAP_ENDURE_RASTA	-1	-1	-1	-1	-1
MOTEADOS_RASTA	No	Si	No	No	No
PROFUND_MOTEADOS_RASTA	-1	18.82	-1	-1	-1
MOTEADOS_MAS70cm._RASTA	No	No	No	No	No
ESTRUCTURA_RASTA	Granular	Granular	Aterronada	Aterronada	Aterronada
OBSERVA_EROSION_RASTA	No	No	No	No	No
OBSERVA_MOHO_RASTA	No	No	No	No	No
OBSERVA_COSTRAS_DURAS_RASTA	No Hay	No Hay	No Hay	No Hay	No Hay
SITIO_EXPUESTO_SOL_RASTA	Mañana y Tarde	Mañana y Tarde	Mañana y Tarde	Mañana y Tarde	Mañana y Tarde
OBSERVA_COSTRAS_BLANCAS_RASTA	No Hay	No Hay	No Hay	No Hay	No Hay
OBSERVA_COSTAS_NEGRAS_RASTA	No Hay	No Hay	No Hay	No Hay	No Hay
REGION_SECA_ARIDA_RASTA	No	No	No	No	No
OBSERVA_RAICES_VIVAS_RASTA	Si	Si	Si	Si	Si
PROFUND_RAICES_VIVAS_RASTA	25.2	22.98	26.26	27.79	25.28
OBSERVA_PLANTAS_PEQUENAS_RASTA	Plantas Normales	Plantas Normales	Plantas Normales	Plantas Normales	Plantas Normales
OBSERVA_HOJARASCA_MO_RASTA	No	No	No	Si	No
SUELO_NEGRO_BLANDO_RASTA	No	No	No	No	No
CUCHILLO_PRIMER_HTE_RASTA	Si	Si	Si	Si	Si
CERCA_RIOS_QUEBRADAS_RASTA	No	No	No	No	No
RECUBRIMIENTO_VEGETAL__SUELO_RASTA	Espaciado	Regular	Bueno	Bueno	Bueno
prof_efectiva	45.92	32.23	44.44	43.86	41.99
d.interno	Bueno	Lento a Muy lento	Bueno	Bueno	BuenoP
drenaje_externo	Lento	Ninguno	Ninguno	Lento	Ninguno
Porc_A	0.54	0.34	0	0	0.51
Porc_Ar	14.23	25.57	41.85	14.82	23.59
Porc_ArA	0.48	1.83	3.78	1.22	0.81
Porc_ArL	0.73	1.28	4.56	1.88	4.22
Porc_FrL	5	6.21	8.46	8.68	16.2
Porc_L	2.29	1.94	2.44	3.38	6.71

Variable	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
Porc_F	11.34	20.26	6.32	8.86	14.9
Porc_FAr	53.82	34.75	27.73	54.23	29.79
Porc_FA	10.82	7.59	4.88	5.96	3.08
Porc_AF	0.75	0.24	0	0.97	0.18
Porc_BLANDO	19.88	61.13	15.9	19.63	17.51
Porc_DURO	11.77	16.36	14.43	10.46	25.86
Porc_EXT_DURO	0.33	0.81	0	0	0
Porc_FRIABLE	45.73	8.97	23.94	33.4	20.92
Porc_FIRME	21.08	9.75	40.94	34.11	31.75
Porc_EXT_FIRME	0.45	0.29	0.22	0	0.44
Porc_PLASTICO	0.76	2.53	3.74	2.39	3.32
Porc_MUY_PLASTICO	0	0.16	0.83	0	0.19
Temp_Max_Avg_Veg	33.55	33.12	33.26	33.17	33.21
Temp_Min_Avg_Veg	24.46	24.41	24.43	24.36	24.47
Temp_Avg_Veg	29	28.77	28.84	28.77	28.84
Diurnal_Range_Avg_Veg	9.1	8.71	8.83	8.81	8.74
Sol_Ener_Accu_Veg	15374	15274	15220	15130	15015
Temp_Max_34_Freq_Veg	0.39	0.3	0.36	0.35	0.33
Rain_Accu_Veg	195.5	221.9	263.4	279.8	212.3
Rain_10_Freq_Veg	0.13	0.16	0.18	0.18	0.15
Rhum_Avg_Veg	80.61	82.42	80.43	81.13	81.2
Temp_Max_Avg_For	33.71	33.26	33.48	33.4	33.57
Temp_Min_Avg_For	24.21	24.14	24.09	24.06	24.29
Temp_Avg_For	28.96	28.7	28.79	28.73	28.93
Diurnal_Range_Avg_For	9.5	9.12	9.4	9.34	9.28
Sol_Ener_Accu_For	22742	22920	22851	22870	22687
Temp_Max_34_Freq_For	0.4	0.29	0.34	0.32	0.35
Rain_Accu_For	268.8	247.2	256.1	233.9	245
Rain_10_Freq_For	0.16	0.14	0.15	0.14	0.14
Rhum_Avg_For	80.77	82.02	79.95	80.22	80.6
Temp_Max_Avg_Mad	33.19	32.7	32.81	32.93	33.01
Temp_Min_Avg_Mad	24.04	23.93	23.54	23.5	23.99
Temp_Avg_Mad	28.62	28.32	28.18	28.22	28.5
Diurnal_Range_Avg_Mad	9.14	8.77	9.26	9.43	9.02
Sol_Ener_Accu_Mad	13762	14223	13897	15296	14544
Temp_Max_34_Freq_Mad	0.31	0.17	0.2	0.23	0.25
Rain_Accu_Mad	117	119.1	138.5	145.5	125.8
Rain_10_Freq_Mad	0.1	0.1	0.13	0.12	0.11
Rhum_Avg_Mad	81.5	81.93	80.62	80.06	80.98
RDT_AJUSTADO	4768	4568	5102	4902	4335

El análisis de los modelos obtenidos permitió identificar diferentes características que influyen en cada uno de los clústeres, como es el caso del número de controles químicos para maleza realizados entre las etapas de siembra y emergencia de las plantas que fue influyente únicamente en uno de los cinco clústeres. Se detectó la importancia de la altura del lote en cuatro de los cinco clústeres y el valor promedio fueron distintos en cada uno de los centroides, los valores de la altura en estos fue de 9.48, 12.48, 11.66 y 12.69 y la correlación de esta variable con el rendimiento es de 0.14, 0.56, 0.11 y -0.12 respectivamente (**Figura 16**). La incidencia de esta variable en los diferentes clústeres es un factor importante en AEPS que ayuda a identificar qué impacta de manera diferente en la producción (rendimiento) de cada clúster.

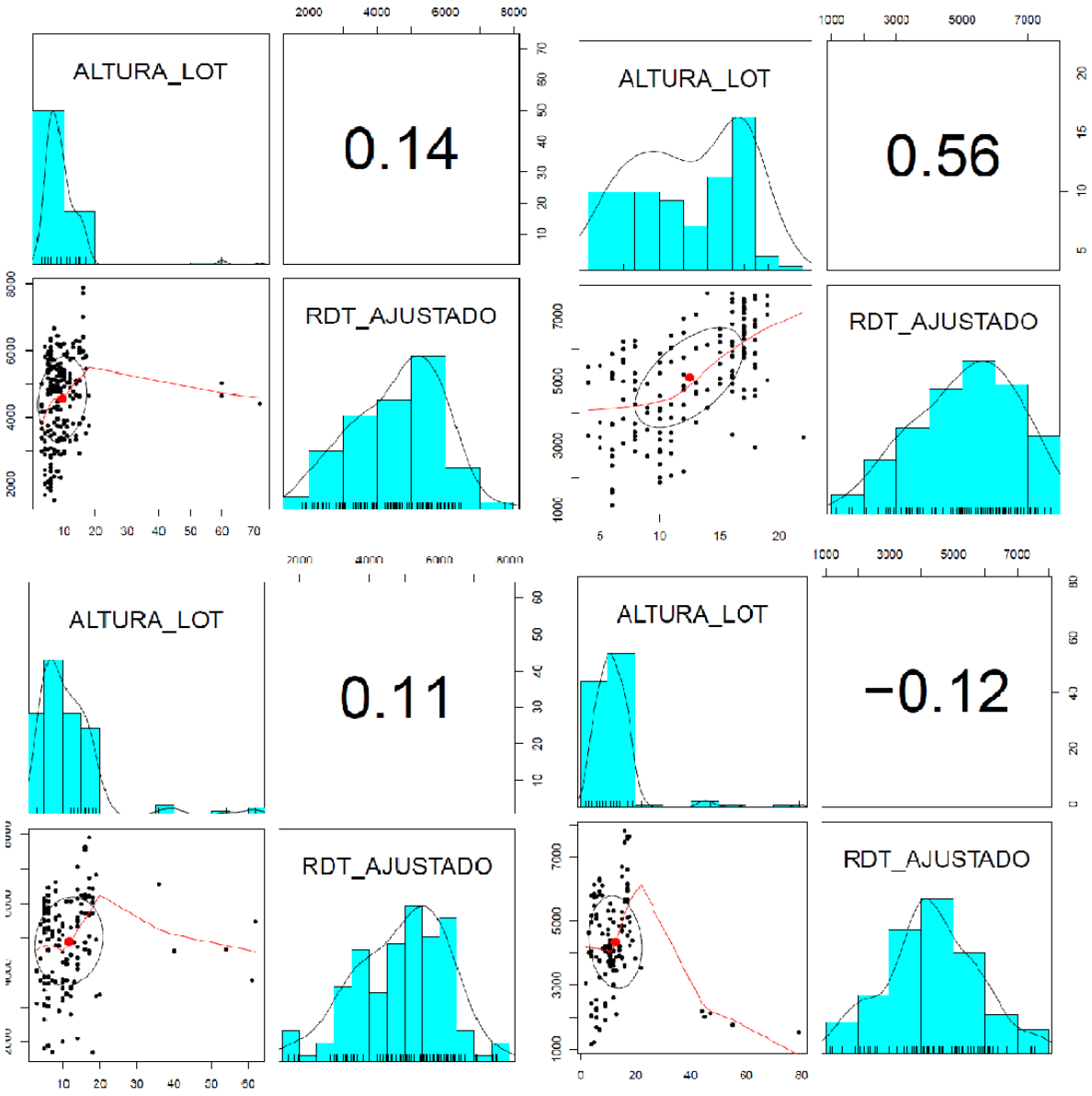


Figura 16. Correlación entre altura y rendimiento por clúster

En los cinco clústeres se encuentran seleccionadas las siguientes variables:

- La semilla utilizada en la siembra (MATERIAL_GENETICO).
- La densidad de plantas a los 20 días (POBLACION_20DIAS_AJT).
- Intensidad de plantas con poca producción (OBSERVA_PLANTAS_PEQUENAS_RASTA).
- El recubrimiento vegetal del suelo (RECUBRIMIENTO_VEGETAL__SUELO_RASTA).
- La presencia de capas duras que pueden impedir el crecimiento de las raíces, el movimiento del agua y la respiración del suelo (CAP_ENDURE_RASTA).

Lo anterior indica que estas variables son relevantes en cualquiera de los escenarios de cultivo identificados, sin embargo, el nivel de correlación entre estas y el impacto en el rendimiento del cultivo varía en cada clúster, con lo que se llega nuevamente a la conclusión que es un factor importante para AEPS la identificación de los diferentes niveles de importancia e incidencia en cada clúster.

Teniendo en cuenta los diferentes modelos de regresión lineal múltiple obtenidos y los coeficientes asociados a las variables, los coeficientes de la variable del material genético denotan que el impacto al interior de cada clúster es diferente, además ayudó a identificar que no todos los materiales proporcionan el mismo efecto si son utilizados en todos los clústeres, que es uno de los factores importantes en AEPS. A continuación, en la **Tabla 13** se indican las mejores opciones del material genético por clúster en donde el agricultor tiene la posibilidad de elegir si su cultivo presenta características similares de suelo y clima a los patrones identificados en los cinco clústeres.

Tabla 13. Material genético por clúster en el mejor macro modelo encontrado

Material Genetico	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
PIONEER 30F35 H	X	X		X	
PIONEER 30F35 HRR	X	X			
ADV 9293 (Syngenta)		X	X		
P3966 (Pioneer)		X		X	
ADV 9339 (Syngenta)				X	X
ICA V 305	X				
PIONEER 30F32	X				
Sinko (Syngenta)	X				
DK7088		X			
PIONEER 30F32HW			X		
Impacto (Syngenta)			X		
DK 1596			X		
Cerato (Syngenta)			X		
P4082 (Pioneer)				X	
Otro[7]				X	
PIONEER 30F35					X
DK7088					X
DK 234					X
PAC 105					X

[7] Material genético no perteneciente al listado de opciones disponibles en el formulario del FENALCE al interior de la plataforma de SIRIA.

Finalmente se logró comprobar que los modelos obtenidos en la solución propuesta usando el R^2 *ajustado* asociado a cada modelo de regresión de los clúster es mayor en un 13% en promedio con respecto al uso de regresión lineal múltiple tomando todas las observaciones como si pertenecieran a un único grupo (dataset total), con lo que se demostró que el enfoque clusterwise aportó de manera significativa al desempeño de los cultivos de maíz y se espera que pueda usarse en otros cultivos, ya que es capaz de detectar distintas relaciones que se dan únicamente en sub conjuntos de la población total y logra el objetivo esperado en la AEPS.

CAPÍTULO 5

5 OPTIMIZACIÓN

En este capítulo se presenta la propuesta de optimización del rendimiento en cultivos de maíz utilizando los modelos construidos bajo el enfoque clusterwise que se presentó en el capítulo 4, posteriormente se presenta la experimentación, los resultados obtenidos, las conclusiones y finalmente el despliegue de la propuesta de optimización en una aplicación web.

En los cultivos de maíz existen diferentes variables que inciden en el rendimiento del cultivo, estas variables se pueden considerar como ***variables fijas*** (que no se pueden controlar o manipular, por ejemplo, las propiedades físico-químicas del suelo y las condiciones climáticas y de temperatura) y ***variables de manejo*** que dependen de diferentes prácticas y cuidados que el agricultor lleva a cabo durante las diferentes etapas del cultivo, como el uso de insumos, sistemas de riego, controles preventivos y correctivos de plagas entre otras. Estas prácticas pueden incidir de forma positiva o negativa sobre el rendimiento, además de ser muy particulares según los conocimientos y experiencia del agricultor.

Teniendo en cuenta lo anterior se eligió un conjunto de 36 variables de manejo pertenecientes a la vista minable definida en el capítulo 3 para realizar la optimización, dichas variables se validaron con expertos del CIAT, las variables están relacionadas con la aplicación de algunos insumos en las diferentes etapas del cultivo de maíz, los controles realizados en las etapas, el tipo de siembra, el método de cosecha y otras como el material genético de las semillas y el número de plantas a los 20 días.

La finalidad de la optimización es poder determinar una combinación de valores en las 36 variables seleccionadas y lograr un efecto positivo sobre el rendimiento en un cultivo de maíz a priori (antes de que se inicie el cultivo).

Como el proceso de optimización es a priori, se requiere definir y obtener diferentes valores de las variables fijas, por ejemplo obtener de servicios externos de predicción, datos del clima y temperatura pronosticados para el periodo del futuro cultivo, además de conocer las características físico-químicas del suelo donde se va realizar el cultivo, todo lo anterior con el fin de poder clasificar el nuevo cultivo en cualquiera de los *k* grupos pertenecientes a los modelos de la solución construida en el capítulo 4 (grupos calculados por el enfoque clusterwise) y llevar a cabo la optimización.

Para la optimización se propone la adaptación de dos metaheurísticas, la mejor búsqueda armónica global (GBHS) y la optimización por enjambre de partículas (PSO), las cuales realizan una búsqueda de la combinación óptima de valores que

deben tomar las 36 variables a optimizar con el objetivo de maximizar el rendimiento.

Las metaheurísticas toman como entrada una solución del proceso de clusterwise para seleccionar el modelo de regresión lineal más apropiado para iniciar el nuevo cultivo, y posteriormente usarlo como parte de la función objetivo en las metaheurísticas para maximizar el rendimiento.

Al finalizar el proceso de optimización se obtiene una solución que corresponde a una observación con los respectivos valores optimizados y el rendimiento acorde al modelo de regresión lineal múltiple aplicado.

5.1 CLASIFICACIÓN INICIAL

La clasificación inicial de un nuevo cultivo se encarga de determinar de la lista de los *k* clúster, cual es el centroide más parecido o cercano en términos de distancia al nuevo cultivo (que va a iniciarse), este proceso se presenta en el **Pseudocódigo 6** y posteriormente se explica en detalle.

```
Entrada:    Nuevo cultivo obs,
            Lista de variables optimización listVariables,
            Solución enfoque listModelos,
Salida:     Modelo clúster más parecido
1   modeloSimilar = nulo
2   minDistancia
3   para i = 0 hasta i < listModel.tamaño hacer
4       auxDistancia = distanciaCluster
                (obs,listModel[i].cluster,listVariables)
5       si modeloSimilar == nulo || auxDistancia < minDistancia entonces
6           minDistancia = auxDistancia
7           modeloSimilar = listModel[i]
8       fin si
9   fin para
10  retornar modeloSimilar
```

Pseudocódigo 6 Clasificación de una nueva observación

Un nuevo evento de cultivo es una observación que puede contener variables de manejo vacías, las cuales se desconocen y corresponden a las variables a optimizar, también está compuesto por las variables fijas las cuales son muy importantes en la clasificación de la observación en un clúster. La clasificación inicial busca encontrar el clúster más cercano haciendo una búsqueda en los clústeres que pertenecen a una solución del enfoque clusterwise.

Para recordar, una solución del enfoque clusterwise (macro modelo) está compuesta por una lista de *k* modelos, cada uno de los cuales tienen en su interior un centroide asociado, estos centroides son los que se tienen en cuenta para determinar cuál es el más parecido a la nueva observación.

Para determinar el clúster más parecido a la nueva observación, se utilizó la distancia euclidiana mixta con una adaptación para excluir las variables a optimizar

las cuales pueden estar vacías y no se deben tener en cuenta. Las únicas variables usadas para el cálculo de distancia son las variables fijas; el clúster que es más parecido a la observación es el que tenga la menor distancia.

En la **Figura 17** se presenta un ejemplo de 3 clústeres que pertenecen a una solución del enfoque clusterwise, estos clústeres están compuestos por observaciones que aparecen en color azul, su respectivo centroide en color amarillo y en color gris la nueva observación que se busca clasificar. Según lo anterior se puede suponer que después de realizar el respectivo cálculo de distancia de cada centroide a la nueva observación, el clúster más parecido es el clúster k1 según los valores de distancia asociados, estos valores aparecen al lado derecho de cada flecha que apunta al respectivo clúster.

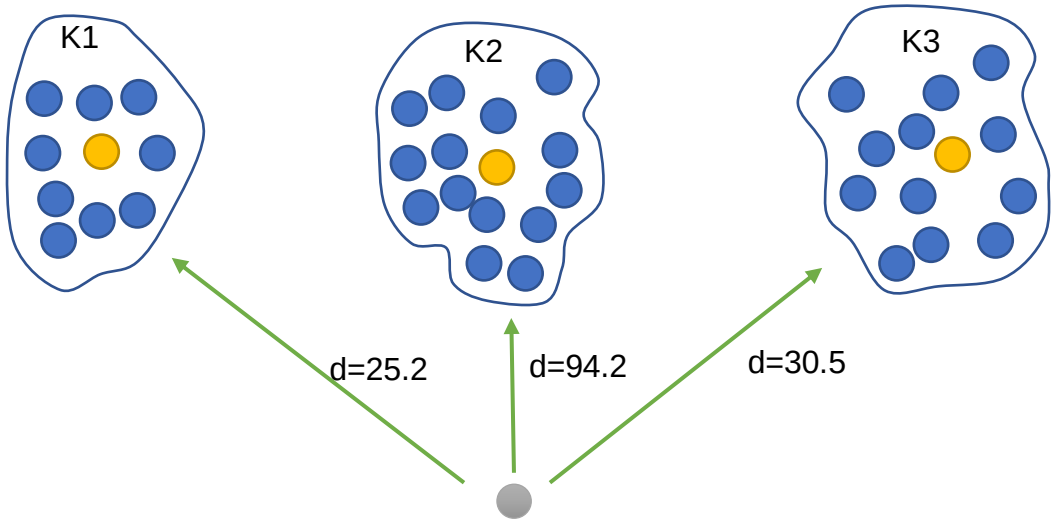


Figura 17. Clasificación de una nueva observación

Finalmente, después de clasificar una observación en un clúster, el modelo de regresión lineal múltiple al cual está asociado ese clúster se utiliza de forma transversal en el proceso de optimización por las metaheurísticas GBHS y PSO propuestas como función objetivo (fitness), las cuales se buscan maximizar el rendimiento.

5.2 APLICACIÓN DEL MODELO DE REGRESIÓN

El modelo de regresión lineal seleccionado en la **sección 5.1**, es el modelo que será usado en la optimización de un cultivo nuevo, recordemos que este modelo fue construido para un clúster conformado por un subconjunto de las 799 observaciones, además tiene como relevantes algunas variables de las 115 variables que inciden en el rendimiento de un cultivo, por tanto todos los modelos de una solución de los *k* clústeres son diferentes.

De acuerdo con lo anterior, como los modelos son diferentes, al aplicar el modelo del clúster al cual fue asignado el nuevo evento de cultivo, quizás algunos de sus valores en las variables nominales no fueron tomados en cuenta en la construcción

del modelo de regresión lineal múltiple, por lo tanto, estos valores se deben omitir debido a que no tienen asociado ningún coeficiente de regresión en el modelo y puede causar errores.

La aplicación del modelo de regresión lineal múltiple se realizó usando un adaptador que fue construido para hacer uso de la librería Weka, este adaptador tiene una función que se presenta en el **Pseudocódigo 7**, en donde se encarga de crear una instancia de Weka a partir de la observación nueva y finalmente usar la función de weka *classifyIntance* para que se aplique el modelo a la nueva observación y esta retornara el valor del rendimiento del cultivo al aplicar el modelo de regresión lineal múltiple.

```
Entrada:    Nuevo cultivo obs,
            Lista de variables a ignorar listVIgnorar,
            Resumen de variables cluster listResumen,
            Dataset data
Salida:     Rendimiento del cultivo aplicando el modelo
1   instancia=crearIntancia(obs, listVIgnorar, listResumen)
2   resultado=modelo.classifyIntance(instancia)
3   retornar resultado
```

Pseudocódigo 7. Aplicación modelo de regresión

Algo muy importante de la clasificación de una instancia en Weka es el proceso necesario para convertir un evento de cultivo nuevo de las estructuras de datos desarrolladas a una instancia de Weka, este proceso se presenta en el **Pseudocódigo 8**.

```
Entrada:    Nuevo cultivo obs,
            Lista de variables a ignorar listVIgnorar,
            Resumen de variables cluster listResumen,
            Dataset data
Salida:     Rendimiento del cultivo aplicando el modelo
1   Instancia = SparceInstance (obs.variables.tamaño-1)
2   instancia.setDataset (dataInstancesModel)
3   j=0
4   para i=0 hasta i< obs.variables.tamaño hacer
5       si listVIgnorar == nulo ||
            listVIgnorar.contiene (i) == falso ||
           (listVIgnorar.contiene (i) &&
           i == data.getIndiceVariableClase()) entonces
6           var = obs.obtenerVariable(i)
7           si var es Continua entonces
8               si var.valor < 0 entonces
9                   instancia.setMissing (j)
10              si no
11                  instancia.setValue (j,var.valor)
12              fin si
13          si no
14              si var es Nominal entonces
15                  resumenNominal = listResumen[i]
16                  si resumenNominal.valores.contiene (var.valor)
                        entonces
```

```
17                                  con=resumenNominal.obtnerValor (i)
18                                  si con >0 entonces
19                                      instancia.setValue (j, var.valor)
20                                  si no
21                                      instancia.setMissing (j)
22                                  fin si
23                              Sino
24                                  instancia.setMissing (j)
25                              fin si
26                          si no
27                              si var es Binaria entonces
28                                  resumenBinario = listResumen[i]
29                                  si (resumenBinario.valor1 == var.valor &&
                                        resumenBinario.countValor1 > 0) ||
                                       (resumenBinario.valor2 == var.valor &&
                                        resumenBinario.countValor2 > 0) entonces
30                                      instancia.setValue (j, var.value)
31                                  si no
32                                      instancia.setMissing (j)
33                                  fin si
34                              si no
35                                  instancia.setMissing (j)
36                              fin si
37                          fin si
38                      fin si
39                  fin si
40          fin para
41          retornar resultado
```

Pseudocódigo 8 crearInstancia

Al construir una instancia se debe tener en cuenta los diferentes casos especiales acorde a los valores que tiene el nuevo cultivo (nueva observación) y el tipo de variable. Para las variables continuas que tienen el valor de -1 este indica que los campos no aplican al cultivo y deben tratarse como atributos vacíos, para las variables nominales y/o binarias si el valor del atributo no existe en el resumen del clúster quiere decir que el valor no fue tenido en cuenta en el modelo y la variable en cuestión debe tomarse como variable vacía para evitar errores, este es el caso cuando existen nuevas prácticas o nuevos materiales genéticos y el modelo no se construyó con esos valores.

5.3 GBHS

La metaheurística de la mejor búsqueda armónica global (GBHS) se utilizó para determinar los valores en las variables de manejo de un nuevo cultivo a optimizar y encontrar el mejor rendimiento en un número de iteraciones dado, tomando como apoyo el modelo de regresión lineal múltiple asociado al clúster más cercano al nuevo evento de cultivo. El funcionamiento de la metaheurística se presenta en el **Pseudocódigo 9**, posteriormente se explica en detalle.

Entrada:	Numero de iteraciones NI, Dataset data, Modelo clúster model Observación a optimizar obs,

```
                Lista de variables a optimizar listVar,
                Lista de variables a ignorar listIgnorar,
                Tasa de consideración memoria armonica HMCR,
                Tamaño de la memoria armónica HMS,
                Ajuste de tono mínimo PAR_MIN,
                Ajuste de tono máximo PAR_MAX
Salida:         Solución optimización
1   memoria = inicializarMemoriaArmonica (model, obs, HMS)
2   para i = 0 hasta i < NI hacer
3       solucion = nulo
4       indiceBest = posicionMejor (memoria)
5       calcularPAR (i, PAR_MIN, PAR_MAX, NI)
6       para j = 0; hasta j < data.variables.tamaño hacer
7           si listVar.contiene (j) || listIgnorar.contiene (j) == falso
                entonces
8               si aleatorio (0, 1) < HMCR entonces
9                   Índice = aleatorio (0, HMS-1)
10                  obs.actualizarVariable (j,
                        memoria[indice].obsevacion.variable (j))
11                  si aleatorio (0, 1) <= PAR entonces
12                      obs.actualizarVariable (j,
                        memoria[indiceBest].obsevacion.variable (j))
13                  fin si
14              si no
15                  obs.actualizarVariable (j, obtenerValor(j))
16              fin si
17          fin si
18      fin para
19      solucion.actualizarObservacion (obs)
20      Solucion.actualizarFitness (evaluarSolucion (model, obs))
21      si existeSolucion (solucion) == false entonces
22          remplazarPeor (solucion)
23          indiceBest = posicionMejor (memoria)
24      fin si
25  fin para
26  retornar memoria[indiceBest]
```

Pseudocódigo 9. GBHS

Una solución de GBHS es una observación con sus 115 variables que tiene asociado un *fitness* (rendimiento del cultivo) el cual es el resultado de la aplicación de un modelo de regresión lineal múltiple asociado al clúster resultado de la clasificación inicial.

La metaheurística tiene unas partes importantes que se pueden resumir en la inicialización de la memoria armónica (línea 1), el proceso de optimización del nuevo cultivo (línea 3 a 20) y actualización de la peor solución (línea 22 y 23) después del proceso iterativo de optimización.

La inicialización de la memoria armónica debe tener en cuenta que se lleva a cabo de manera parcial tomando como base el evento de cultivo que tiene unas variables de suelo, clima y otras que son fijas, además una lista de variables a optimizar que corresponden a las variables de manejo que son las susceptibles de cambios basados en el clúster al que fue asignado en la clasificación inicial.

La inicialización de la memoria armónica consiste en la creación de **HMS** nuevas soluciones a partir de los valores de la observación **obs** (nuevo cultivo) y los rangos de datos de las variables del clúster en la que esta observación fue clasificada. Esta creación de las nuevas observaciones se realizó siempre y cuando la combinación de los valores posibles de las variables a optimizar de lugar a más de **HMS+NI** soluciones diferentes, de lo contrario, si la(s) variable(s) a optimizar son solo nominales y/o binarias y las combinaciones de estas no superen el tamaño de la memoria armónica (**HMS**), se realiza la optimización de una única observación de manera exhaustiva con todas las posibles combinaciones y se retornara la mejor solución (mejor rendimiento). Si dentro de las variables a optimizar existen variables continua esta dará lugar a infinitas combinaciones en conjunto con otras (nominales y/o binarias), por ende, se ejecuta el proceso de optimización con normalidad.

Las soluciones u observaciones de la memoria armónica se construyen a partir de una clonación de la observación **obs** y de un proceso de elección aleatorio de las variables a optimizar, usando la información del rango de valores de las variables del clúster del modelo según la solución del enfoque clusterwise seleccionado para la observación. El proceso de selección de los valores se hace de forma aleatoria tomando en cuenta la estructura de control del clúster, en donde según el tipo de variable a optimizar se hace la elección, para las variables continuas se genera un número aleatorio según el rango de cada variable, para las variables nominales y binarias se elige de forma aleatoria un valor dentro de los posibles valores de la variable en el clúster. Es bueno resaltar que tras la creación de una nueva observación se verifica que sea única, es decir que no esté en la memoria armónica para poderse agregar sin incurrir en problemas de convergencia prematura, este proceso se lleva acabo comparando solamente las variables que se están optimizando ya que el resto de las variables son iguales. Para las variables continuas, la comparación se tiene en cuenta como igual si la diferencia entre ambos valores a comparar es inferior a $1e^{-5}$ debido a que si la diferencia es inferior esta no tendrá un efecto significativo al aplicar el modelo de regresión lineal múltiple, es decir se obtienen los mismos resultados para 5.15689 y 5.15688 a pesar de ser diferentes.

Después de construir las observaciones de la memoria armónica se ubica la posición en la memoria de la mejor (mayor valor de fitness) y la peor solución (menor valor de fitness), teniendo en cuenta la evaluación de la función objetivo corresponde a la aplicación del modelo de regresión lineal múltiple a la observación. La evaluación obtiene el valor del rendimiento de un cultivo.

El proceso de optimización de un nuevo cultivo se ejecuta repetidamente durante **NI** (número de iteraciones) iteraciones (bucle ***para*** líneas 2 a 25) buscando maximizar el *fitness*. Durante estas iteraciones se crea una solución a optimizar con las variables fijas iguales a la mejor solución de la memoria y en la cual se definen posteriormente las variables de manejo según las 3 reglas de improvisación del algoritmo GBHS, así:

- Primero (líneas 8 a 13), se genera un número aleatorio entre 0 y 1 y si este número es inferior al parámetro **HMCR** (tasa de consideración de la memoria armónica) entonces se elige de forma aleatoriá una observación de la memoria armónica, y se actualiza la variable de la observación a optimizar por el valor que tiene la observación elegida en la memoria armónica. Después se genera un nuevo número aleatorio entre 0 y 1, y si esté número es menor o igual al parámetro ***PAR*** (tasa de ajuste del tono), entonces nuevamente se actualiza la variable que se está optimizando de la observación con el valor de la variable de la mejor solución. Es bueno resaltar que el parámetro ***PAR,*** es dinámico y por eso se calcula para cada iteración (línea 5) según la ecuación 8.

$$PAR = PAR_{MIN} + \left(\frac{PAR_{MAX} - PAR_{MIN}}{NI - 1}\right) * (i - 1) \qquad 8$$

Donde ***PAR_MIN*** es el mínimo valor de la tasa de ajuste de tono definido por el usuario***, PAR_MAX*** es el máximo valor de la tasa de ajuste de tono definido por el usuario, ***NI*** es el número máximo de iteraciones e ***i*** es el número de la iteración actual.

- Segundo (líneas 14 a 16), en caso de que el número aleatorio generado no supere a **HMCR** entonces se actualiza el valor de la variable a optimizar por un valor tomado de forma aleatoria del clúster según la estructura de control en los rangos permitidos y de acuerdo con el tipo de variable (continua, categórica o binaria).

Después de realizar la definición de todas las variables, se actualiza el *fitness* de la observación optimizada, ejecutando una evaluación del modelo de regresión lineal múltiple (línea 20). Posteriormente se verifica que la observación optimizada no esté en la memoria armónica (línea 21), de ser afirmativo se compara esta con la peor solución, si la observación optimizada supera el *fitness* de la peor se ingresa esta observación optimizada en la memoria armónica en remplazo de la peor y se actualizan las posiciones de la mejor y peor solución de la memoria armónica, esto con el objetivo de poder realizar siempre una optimización sobre la mejor solución y poder guardar en la lista las mejores soluciones, en caso que la observación optimizada esté en la memoria armónica se descarta.

5.4 PSO

La metaheurística de optimización por enjambre de partículas (PSO) también se utilizó para determinar los valores en las variables de manejo de un nuevo cultivo a optimizar y encontrar el mejor rendimiento en un número de iteraciones dado, tomando como apoyo el modelo de regresión lineal múltiple asociado al clúster más cercano al nuevo evento de cultivo. El funcionamiento de la metaheurística se presenta en el **Pseudocódigo 10** y posteriormente se explica en detalle.

Entrada:	Numero de iteraciones NI, Numero de partículas NP,

```
                Radio de velocidad RV,
                ALPHA,
                GAMMA,
                DELTA,
                BETA,
                Dataset data,
                Modelo cluster model
                Observación a optimizar obs,
                Lista de variables a optimizar listVar,
                Lista de variables a ignorar listIgnore
Salida:         Solución optimización
1   particulaBest = nulo
2   para i = 0 hasta i < NP hacer
3       enjambre [i] = nuevaParticula (padre)
4   fin para
5   para i = 0 hasta i < NI hacer
6       para j = 0 hasta j < NP hacer
7           Si (particulaBest == nulo ||
                enjambre[j].sBest.fitness > particulaBest.sx.fitness)
                entonces
8               particulaBest = enjambre[j]
9           fin si
10      fin para
11      para j = 0 hasta j < NP hacer
12          enjambre[j].actualizarVelocidad (particulaBest)
13      fin para
14      para j = 0 hasta j < NP hacer
15          enjambre[j].actualizarSx ()
16      fin para
17  fin para
18  retornar particulaBest.best
```

Pseudocódigo 10. PSO

Al igual que GBHS, PSO tiene como solución la misma observación con sus 115 variables y un *fitness* (rendimiento del cultivo) obtenido de la aplicación del modelo de regresión lineal múltiple.

La metaheurística tiene una población inicial que corresponde a **NP** (número de partículas) soluciones, las cuales se encargan de realizar una exploración y explotación del espacio de búsqueda para encontrar una solución que maximice su *fitness*. La búsqueda de mejores soluciones se logra a partir del conocimiento colectivo de la mejor partícula global del enjambre y la memoria local que tiene cada partícula de la mejor solución que ha encontrado a partir de su propio proceso de búsqueda.

Existen cuatro partes importantes de PSO las cuales corresponden a la inicialización del enjambre de partículas (líneas 2 a 4), definición de la mejor solución encontrada por las partículas del enjambre (líneas 6 a 10), actualización de los componentes de velocidad de las partículas (líneas 11 a 13) y finalmente la actualización de la posición en el espacio de búsqueda de las partículas (líneas 14 a 16).

La inicialización del enjambre de partículas se encarga de crear un número de partículas (**NP**) las cuales contienen en su interior dos soluciones una solución actual (**Sx**) y la mejor solución que han encontrado hasta el momento (**Best**) de ejecución del algoritmo (al inicio es una copia de Sx), además un enlace al enjambre (padre). Cuando se crean las partículas, las dos soluciones toman como base la observación **obs** (nuevo cultivo), en donde al igual que GBHS tendrá las variables de suelo, clima y otras fijas (que no se pueden modificar) y otras variables a optimizar que corresponden a las variables de manejo.

Teniendo en cuenta lo anterior, al crear una partícula, los diferentes valores de las variables a optimizar se eligen de forma aleatoria tomando información de la estructura de control del clúster más parecido al cual se clasificó inicialmente la observación del nuevo cultivo, además después de tener una solución inicial esta se evalúa (se le define el fitness) aplicando el modelo de regresión lineal múltiple y esta pasa a ser **Sx** y **Best**. Es bueno resaltar que cada partícula tiene además un componente de velocidad por cada variable, el cual se inicializa con el radio de velocidad (parámetro del algoritmo) del enjambre.

Después de la creación de la población inicial se comienza un proceso iterativo (líneas 5 a 17) que se ejecuta durante un máximo número de iteraciones (**NI**), lo segundo que se hace es determinar al comienzo de cada iteración, cual es la mejor partícula, la cual corresponde a la que tenga la mejor solución es decir mayor *fitness*, por tanto se recorre el enjambre y se verifican las soluciones **Best** de cada partícula y se elige la partícula con el mejor *fitness*.

Lo siguiente es la actualización de los componentes de velocidad de las partículas. Es preciso tener en cuenta que cada partícula tiene un vector de velocidad con 115 componentes, los cuales están asociados a cada variable. La velocidad en PSO ayuda a orientar el desplazamiento de la partícula por el espacio, si el valor del componente es pequeño se desplaza lentamente permitiendo hacer más explotación y si el componente es grande se desplaza rápidamente haciendo más exploración.

Como los componentes de velocidad intervienen en el desplazamiento, fue necesario realizar una adaptación en la actualización de estos para que funcione para variables mixtas, debido a que no solo se cuenta con variables continuas, sino también con variables nominales (categóricas) y binarias, las cuales se van a optimizar. En el **Pseudocódigo 11** se presenta como se hace la actualización de los componentes de velocidad de cada partícula solamente para las variables de tipo continuo y binario.

Entrada: Mejor partícula global pGlobal,
Enjambre padre
Salida:

```
1    b = d = 0
2    para j = 0 hasta padre.listVar.tamaño hacer
3        I = padre.listVar[j]
4        varSx = Sx.obtenerVariable(i)
```

```
5           b = padre.aleatorio (0, padre.GAMMA)
6           d = padre.aleatorio (0, padre.DELTA)
7           alea1 = padre.aleatorio (0, 1)
8           alea2 = padre.aleatorio (0, 1)
9           auxVel = 0
10          varBest = Best.obtenerVariable (i)
11          varGlobal = pGlobal.Best.obtenerVariable (i)
12          si varSx es Continua entonces
13                  auxVel = alpha * velocidad[i] +
                            padre.GAMMA * alea1 * (varBest.valor - varSx.valor) +
                            padre.DELTA * alea2 * (varGlobal.valor - varSx.valor)
14                  velocidad[i] = auxVel
15                  si velocidad[i] < -padre.radioVelocidad entonces
16                          velocidad[i] = -padre.radioVelocidad
17                  si no
18                          si velocidad[i] > padre.radioVelocidad entonces
19                                  velocidad[i] = padre.radioVelocidad
20                          fin si
21                  fin si
22          si no
23                  si varSx es Binaria entonces
24                          auxVel = Alpha * velocidad[i] +
                                    padreGAMMA * alea1 * (varBest.valor == varSx.valor?0:1) +
                                    padre.DELTA * alea2 * (varGlobal.valor == varSx.valor?0:1)
25                          velocidad[i] = auxVel
26                  fin si
27          fin si
28          alpha=alpha*padre.BETA
29  fin para
```

Pseudocódigo 11. Actualizar velocidad

La actualización de cada uno de los componentes de velocidad se hace de acuerdo al tipo de variable y se presenta en las líneas 13 y 24 del **Pseudocódigo 11**, en donde la actualización de cada componente interviene el valor actual de velocidad e información de los valores de la variable relacionada al componente de la mejor solución **Best**, la solución actual **Sx** y la mejor solución global **pGlobal**. Además, se resalta que cuando una variable es binaria no se hace la diferencia de componentes como en una variable continua, sino que se comparan los valores que tienen, si son iguales se asigna 0 de lo contrario se asigna 1. Como se puede apreciar, las variables nominales (categóricas) no hacen uso de los valores de velocidad y por eso no se actualizan.

Después de actualizar los componentes de las variables a optimizar, sigue la parte de actualización de la solución actual de las partículas del enjambre. Teniendo en cuenta que la actualización de la solución se realiza con base en el componente de velocidad asociado a la variable, se realizó un ajuste para las variables binarias y nominales, buscando que el nuevo valor estuviera en el rango correcto de la variable según la estructura de control. El **Pseudocódigo 12** presenta el proceso de actualización de la solución **Sx** en una partícula.

Entrada: Enjambre padre

Salida:

```
para j = 0 hasta padre.listVar.tamaño hacer
    I = padre.listVar[j]
    varSx = Sx.obtenerVariable (i)
    si varSx es Continua entonces
        aux = varSx.valor + padre.e * velocidad[i]
        si aux >= padre.model.resumenCluster.variable[i].min &&
            aux <= padre.model.resumenCluster.variable[i].max entonces
            Sx. actualizarVariable(i,aux)
        si no
            si aux < padre.model.resumenCluster.variable[i].min entonces
                Sx.actualizarVariable (i,
                    padre.model.resumenCluster.variable[i].min + velocidad[i])
            Sino
                Sx. actualizarVariable (i,
                    padre.model.resumenCluster.variable[i].max - velocidad[i])
            fin si
        fin si
    si no
        si varSx es Binaria entonces
            sigVelocidad =1 / (1 + elevar(e, (-1 * velocidad[i])))
            alea = padre.aleatorio (0,1)
            resumenBinario = c
            valor = alea <sigVelocidad? resumenBinario.valor1:
                                        resumenBinario.valor2
            Sx.actualizarVariable (i, valor)
        si no
            si varSx es Nominal entonces
                varBest = Best.obtenerVariable(i)
                varGlobal = padre.particulaBest.Best.obtenerVariable (i)
                alea1 = padre.aleatorio(0,1)
                valorNominal = nulo
                si alea1< alpha entonces
                    valorNominal = seleccionarValorDiferente (i, varSx.valor)
                si no
                    valorNominal = varSx.valor
                fin si
                alea2 = padre.aleatorio (0, 1)
                si alea2 > padre.GAMMA entonces
                    valorNominal = varBest.valor
                fin si
                alea3 = padre.aleatorio (0, 1)
                si alea3 > padre.DELTA entonces
                    valorNominal = varGlobal.valor
                fin si
                Sx.actualizarVariable (i, valorNominal)
            fin si
        fin si
    fin si
fin para
evaluarParticula ()
```

Pseudocódigo 12. Actualizar solución actual Sx

Para las variables continuas la velocidad tiene una incidencia directa sobre el valor que va a tomar la posición de la solución **Sx** en la variable *i*, en donde al valor de la

variable *i* se le incrementa el valor de velocidad asociada, pero si al realizar este incremento hace que la variable i se salga del rango entonces la variable comienza en alguno de los extremos inferior o superior y según sea el caso se incrementa o decrementa la velocidad (líneas 4 a 14).

Para las variables binarias (líneas 16 a 21) la elección del valor a tomar en la posición de la solución **Sx** en la variable *i* se hace por medio de la evaluación de función sigmoidal del valor de la velocidad (es decir la velocidad se toma como una probabilidad de que el valor de posición sea cero o uno). El comportamiento de esta función se presenta en la **Figura 18**.

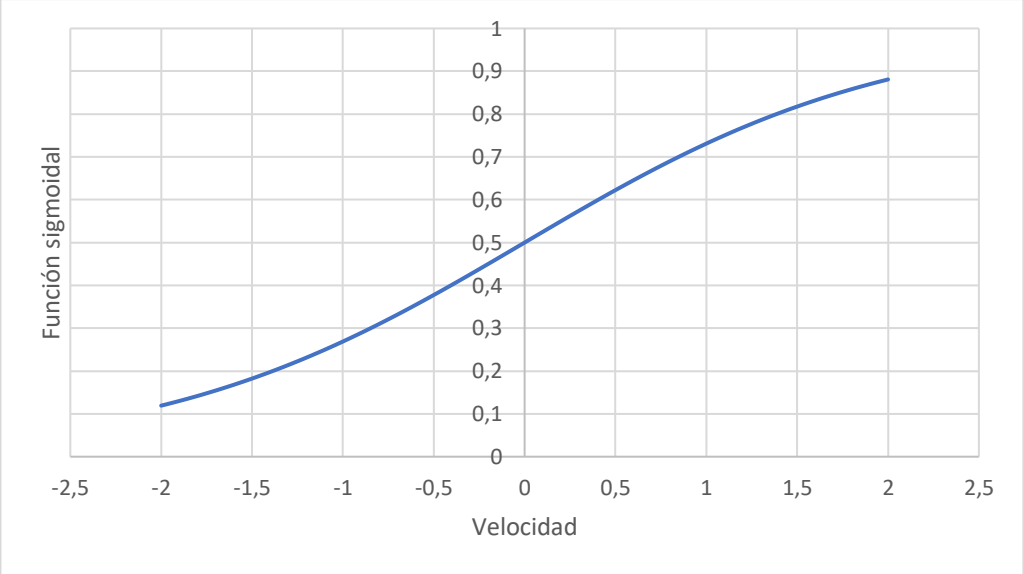


Figura 18. Comportamiento función sigmoidal

El uso de la función sigmoidal se hace con el objetivo de elegir cual es el valor que va a tomar la variable binaria de los dos posibles, usando la velocidad asociada, para ello se genera un numero aleatorio de 0 a 1 y si este es inferior a la evaluación de la velocidad en la función sigmoidal toma el valor 1 de la variable de lo contrario se le asigna el valor 2. Si la variable fue seleccionada en el modelo, esto quiere decir, que en la estructura de control del clúster en la variable binaria si solo existe un valor no se aplicará este proceso a la variable y siempre tomará el mismo valor. Por ejemplo, suponga que la variable sistema de riego en el cultivo puede ser SI o NO, pero al realizar la clasificación inicial del nuevo cultivo es similar al clúster 1 y este en la variable sistema de riego está compuesto solo por cultivos que tiene el valor SI, por tanto no aplicaría optimización para esta variable porque solo podrá tomar el valor SI.

Para las variables nominales (categóricas), como no interviene el componente de velocidad, la actualización del valor de la posición de la variable *i* en la solución **Sx** se hace utilizando probabilidades, aprovechando la información de la mejor solución **Best** y la mejor solución global. El proceso se hace en tres pasos:

- Paso 1: Se genera un número aleatorio entre 0 y 1, si este es menor el valor del parámetro **alpha** el valor que toma la variable *i* es un valor aleatorio de los posibles que pueda tomar diferente al que tiene asignado actualmente, en caso contrario que el valor aleatorio sea mayor que el parámetro **alpha** se dejara el mismo valor que tiene.
- Paso 2: Se genera un número aleatorio entre 0 y 1, si este es mayor que el valor del parámetro **GAMMA** del enjambre, el valor que toma la variable *i* es el que tiene la variable *i* de la mejor solución de la partícula **Best**.
- Paso 3: Se genera un número aleatorio entre 0 y 1, si este es mayor que el valor del parámetro **DELTA** del enjambre el valor que toma la variable *i* es el que tiene la variable *i* de la mejor solución global del enjambre.

Finalmente, después de realizar las partes importantes antes descritas, estas se ejecutan hasta cumplir el criterio de parada y se retornara la mejor solución encontrada, que es la solución global del enjambre de partículas.

5.5 APORTES

En la **Tabla 14** se presenta en resumen los aportes y modificaciones hechas en los algoritmos GBHS y PSO.

Tabla 14. Aportes optimización

Tema	**Aporte**
GBHS	La función objetivo se adaptó maximizar el R_2 ajustado de un modelo de regresión lineal múltiple el cual se eligió de un proceso de clasificación inicial. En la clasificación inicial se obtuvo un clúster que tiene el modelo y un centroide con una estructura de control que usara GBHS para optimización. Para la construcción de soluciones de la memoria armónica se hizo una construcción parcial basándose en una observación inicial que tendría valores fijos y un array de las variables a optimizar. Los valores que tomarían las variables a optimizar fueron elegidos de la siguiente forma: • Para las variables continuas se eligió un valor aleatorio dentro de un rango según el rango de la

	estructura de control del centroide para dicha variable. • Para las variables nominales y binarias se eligió un valor aleatorio de una lista de los posibles que pudiese tomar de la estructura de control para dicha variable. En el proceso de construcción de una nueva partitura al tener diferentes estrategias de elección de qué valor tomar teniendo en cuenta probabilidades si tomar de la memoria armónica de cualquier solución, de la mejor solución o un valor aleatorio para el caso de elegir un valor de la mejor solución se cambió teniendo en cuenta la particularidad de las variables.
PSO	PSO se adaptó para trabajar con datos de tipo mixto. La función objetivo se adaptó al igual que GBHS. Para la construcción de soluciones del enjambre se tuvo en cuenta los mismos casos que en GBHS. En el proceso de actualización de la velocidad tratándose de que los datos son mixtos se hizo variación del cálculo de velocidad para las variables binarias y para las variables nominales no se utilizó. En el proceso de actualización de posiciones se aplicó de forma selectiva a las variables a optimizar y se hizo según el tipo de variable: • Para las variables continuas se utilizó la velocidad como componente que define el desplazamiento.

	• Para las variables binarias se hizo una evaluación de la velocidad con una función sigmoidal y acorde a esta evaluación y una probabilidad se define el valor a tomar. Para las variables nominales se optó por usar algunas probabilidades y los valores a tomar dependen de estas.

5.6 EXPERIMENTACIÓN

En este ítem se presenta los parámetros utilizados para las pruebas realizadas con las metaheurísticas GBHS y PSO en el proceso de optimización del rendimiento de los cultivos usando la solución construida por el enfoque clusterwise presentado en el capítulo 4.

La ejecución de los algoritmos de optimización se realizó en un equipo con las siguientes características: procesador Intel Core i7 a 2.4, 8GB RAM, SSD Kingston uv400 y sistema operativo Windows 10.

Durante el proceso de experimentación se ejecutaron pruebas con las 799 observaciones y con una lista de 36 variables a optimizar definidas como variables de manejo con el objetivo de poder contrastar el rendimiento de cultivos existentes con los resultados de optimización. El macro modelo utilizado fue el elegido como la mejor solución obtenida en las pruebas del enfoque clusterwise con 5 clúster. En la **Tabla 15** se presentan las variables elegidas para optimización.

Tabla 15. Variables de manejo optimización

TIPO_SIEMBRA	ContMalQui_Flor_Cose
SEM_TRATADAS	ContPlaQui_Antes_Siem
MATERIAL_GENETICO	ContPlaQui_Siem_Emer
DRENAJE	ContPlaQui_Emer_Flor
METODO_COSECHA	ContPlaQui_Flor_Cose
ALMACENAMIENTO_FINCA	TotN_Antes_Siem
DIAS_EN_EMERGER	TotN_Siem_Emer
DIAS_EN_EMERGER_A_FLORECER	TotN_Emer_Flor
DIAS_EN_FLORECER_A_COSECHAR	TotP_Antes_Siem
POBLACION_20DIAS_AJT	TotP_Siem_Emer
ContEnfQui_Emer_Flor	TotP_Emer_Flor
ContEnfQui_Flor_Cose	TotK_Antes_Siem
ContMalMec_Siem_Emer	TotK_Siem_Emer
ContMalMec_Emer_Flor	TotK_Emer_Flor
ContMalMec_Flor_Cose	FerOrg_Emer_Flor
ContMalQui_Antes_Siem	FerQui_Antes_Siem
ContMalQui_Siem_Emer	FerQui_Siem_Emer

ContMalQui_Emer_Flor	FerQui_Emer_Flor

5.6.1 Afinamiento del algoritmo GBHS

El número de iteraciones elegido fue de 200 siendo el número mínimo de iteraciones donde se encontraron buenas soluciones, el tamaño de la memoria armónica elegido fue de 20, también se probó con otros valores como 10 y 15. La tasa de consideración de la memoria armónica se fijó en 0.85 con el objetivo de que generen el 15% de las variables en forma aleatoria y se haga una apropiada exploración, se probó con un valor inferior pero no dio buenos resultados. El resumen de los parámetros usados se presenta a continuación (es preciso decir, que como trabajo futuro se debe realizar un proceso sistemático de afinación de estos parámetros):

Tasa de consideración de la memoria armónica	HMCR = 0.85
Tamaño memoria armónica	HMS = 20
Tasa mínima ajuste de tono	PAR_MIN = 0.05
Tasa máxima ajuste de tono	PAR_MAX = 0.45
Máximo número de iteraciones	NI = 200
Lista de variables a ignorar	listIgno = [RDT_AJUSTADO, ID_LOTE]
Máximo número de evaluaciones de la función objetivo	220

5.6.2 Afinamiento del algoritmo PSO

Los parámetros definidos para PSO corresponden a recomendaciones de la literatura. El resumen de los parámetros usados se presenta a continuación y también se recomienda como trabajo futuro realizar un proceso sistemático de afinación de estos parámetros:

Numero de iteraciones	NI = 30
Numero de partículas	NP = 10
Radio de velocidad	RV = 2
ALPHA	ALPHA = 0.9
GAMMA	GAMMA = 2
DELTA	DELTA = 2
BETA	BETA = 0.975
Lista de variables a ignorar	listIgno = [RDT_AJUSTADO, ID_LOTE]
Máximo número de evaluaciones de la función objetivo	300 (superior en 80 frente a GBHS)

5.7 COMPARACIÓN DE RESULTADOS

Con las adaptaciones realizadas a las metaheurísticas y el macro modelo seleccionado por su buen fitness y simplicidad, se ingresaron una por una las 799 observaciones a GBHS y PSO para que fueran optimizadas todas sus variables de manejo y con esto encontrar el mejor rendimiento posible.

Ambas metaheurísticas lograron mejorar el rendimiento en la mayoría de las 799 observaciones (**ANEXO 13**); GBHS optimizó 798 observaciones y PSO 782, pero GBHS en su única observación sin optimizar, no se alejó demasiado del valor original, en cambio PSO no logro encontrar un rendimiento mayor a la media tonelada para esta misma observación cuyo rendimiento original era de 4.4 toneladas.

La adaptación realizada a GBHS logró encontrar el mayor valor de rendimiento con 20 toneladas por hectárea (5 más que la máxima reportada por PSO), adicionalmente el promedio de los rendimientos generados es mayor también para GBHS (ver **Figura 19**).

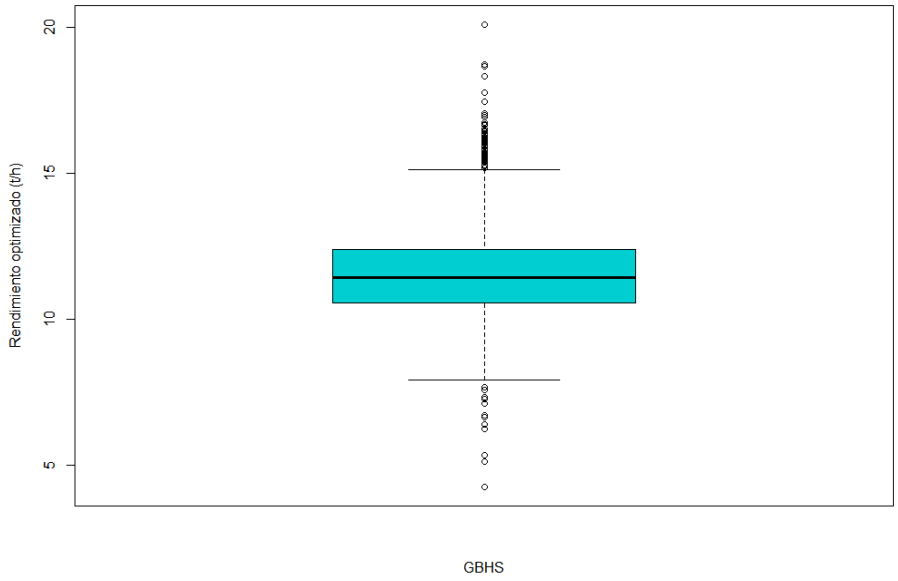


Figura 19. Rendimientos generados por GBHS

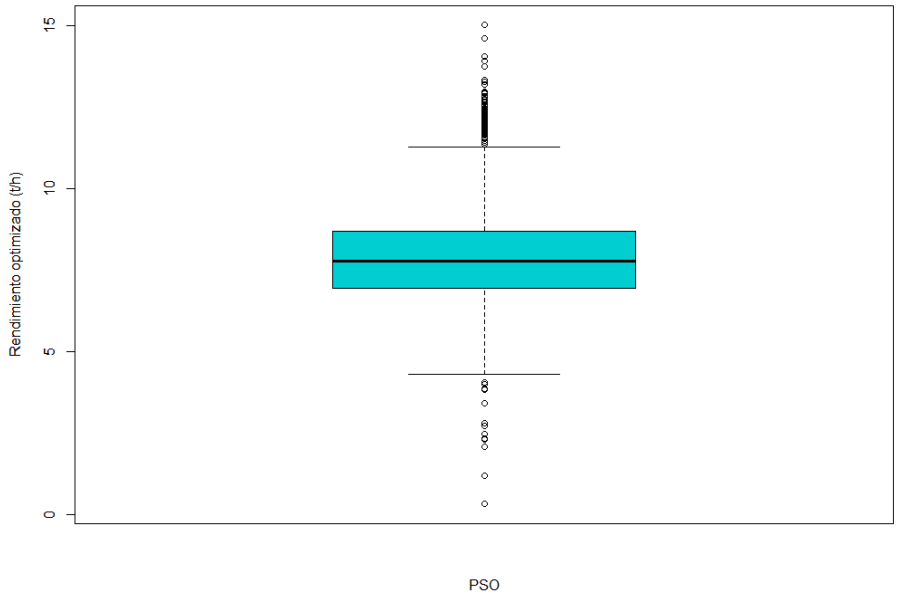


Figura 20. Rendimientos generados por PSO

Los promedios de rendimientos optimizados que se obtuvieron son mayores para GBHS en más del doble frente a PSO, adicionalmente el mayor porcentaje de optimización encontrada se construyó con GBHS, lo anterior se puede observar en la **Tabla 16**.

Tabla 16. Porcentajes de optimizaciones

OPTIMIZACIÓN	GBHS	PSO
MÍNIMA	-3.661%	-92.59%
PROMEDIO	171.93%	84.56%
MÁXIMA	1122.85%	717.24%

Por último, los tiempos de procesamiento también son favorables para GBHS, debido a que en el procesamiento de las 799 observaciones no superó los 2 minutos de ejecución, en su lugar PSO toma un mínimo de 10 minutos para procesar estas mismas observaciones.

5.8 CONCLUSIONES

Existe un evento de cultivo que no fue posible optimizar por ambas metaheurísticas, lo que implica que es posible que tuviera ya definidas las mejores prácticas de manejo y por esta razón no fuese encontrado un mayor rendimiento por parte de GBHS y PSO.

La adaptación de la metaheurística GBHS presentó un mejor desempeño frente a PSO, teniendo en cuenta que GHBS tiene:

- Un menor número de observaciones sin optimizar.
- Un promedio de porcentaje de optimización mayor en más del doble.
- Inferiores tiempos de procesamiento.
- El máximo valor de rendimientos en toneladas reportado.
- Un mayor promedio en rendimientos.

La **Figura 21** permite observar el rendimiento optimizado (Eje x) frente a lo que se encuentra actualmente registrado en las 799 observaciones del dataset (Eje y). Es evidente la mejora lograda. Además, los tiempos de ejecución también son un factor significativo debido a que esta optimización se espera sea utilizada en tiempo real por agricultores mediante un formulario web; al tener la capacidad de brindar respuesta en tiempos menores a una décima parte de segundo facilita su uso debido a que elimina largas esperas o necesidades de realizar el proceso en más de un paso.

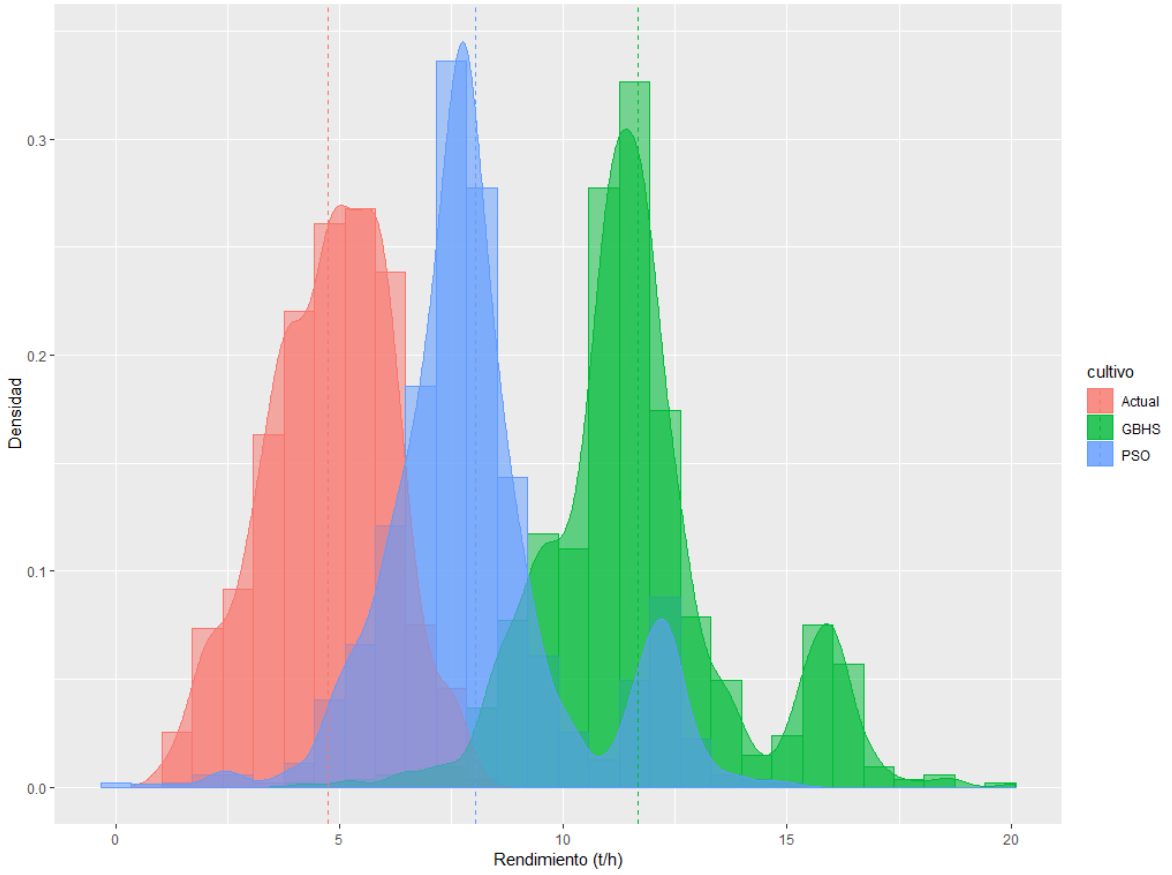


Figura 21. Rendimiento real vs optimizado por PSO y GBHS

Por todo lo anterior se seleccionó la adaptación realizada a la metaheurística de la mejor búsqueda armónica global como la responsable de optimizar el rendimiento en futuros eventos de cultivos de maíz al interior del departamento de Córdoba. También es preciso comentar que en un trabajo futuro se deben incluir restricciones relacionadas con la disponibilidad de recursos económicos (dinero disponible) por parte de los agricultores para que las recomendaciones entregadas sean más aplicables en el contexto real de ellos.

5.9 DESPLIEGUE

Con la selección de GBHS como optimizador por su desempeño superior frente a PSO, se procedió a la creación de un aplicativo web que permite acceder a la optimización de prácticas de manejo agronómicas.

Para la ejecución de esta actividad se desarrolló apoyándose en la metodología SCRUM. Se definió el product backlog, el cual contenía los requerimientos que debería cumplir el aplicativo web, con esto se priorizaron por dependencia e importancia las actividades y estas posteriormente se desarrollaron en dos sprints,

en estos se detalló las actividades a realizar y posteriormente se ejecutaron dichas actividades.

5.9.1 Product backlog

El cliente final (CIAT y FENALCE) fue representado por el profesional Hugo Andrés Dorado Betancourt (director del proyecto), con quien se realizó reuniones para entender el producto que se debería desarrollar. A partir de estas reuniones se establecieron los siguientes requerimientos:

- El aplicativo web contendrá un formulario para el ingreso de la información.
- El formulario debe tener apariencia similar a la plataforma **http://siria.fenalce.org/** que es actualmente utilizada para ingresar la información de eventos de cultivos e información relacionada.
- El formulario debe permitir ingresar la planeación en cuanto a prácticas de manejo relacionadas a un evento de cultivo, estas deben limitarse a las prácticas que terminaron seleccionadas al interior de la vista minable definida en el capítulo 3.
- El formulario debe permitir visualizar las variables relacionadas con la guía RASTA y seleccionadas en la vista minable (esta información a futuro será recuperada desde las APIs propias de FENALCE o quien lo proporcione).
- La aplicación debe permitir cargar eventos de cultivo junto con su respectiva descripción de terreno y clima, esto entendiendo que esta por fuera del alcance del presento proyecto la predicción a futuro del clima y la conexión con aplicaciones de terceros para recuperar información de caracterización de suelos.
- La aplicación web debe enviar la información suministrada al proceso de optimización y retornar las recomendaciones junto con el rendimiento pronosticado.

5.9.2 Sprints

Con los requerimientos definidos por el cliente se procedió a desarrollar estos mediante dos sprints de 2 semanas cada uno.

5.9.2.1 Sprint 1

El primer sprint consistió en definir cómo y en que se realizaría los requerimientos, para lograr esto se desarrollaron las siguientes actividades:

- Conocer la plataforma **http://siria.fenalce.org/** e identificar aspectos visuales a ser utilizados en la aplicación web de optimización, para esto se navegó por la opción de probar el sistema la cual permite hacer uso de la aplicación sin crear una cuenta y trabajar con registros de eventos de cultivos y caracterización de suelos como se observa en la **Figura 22**.

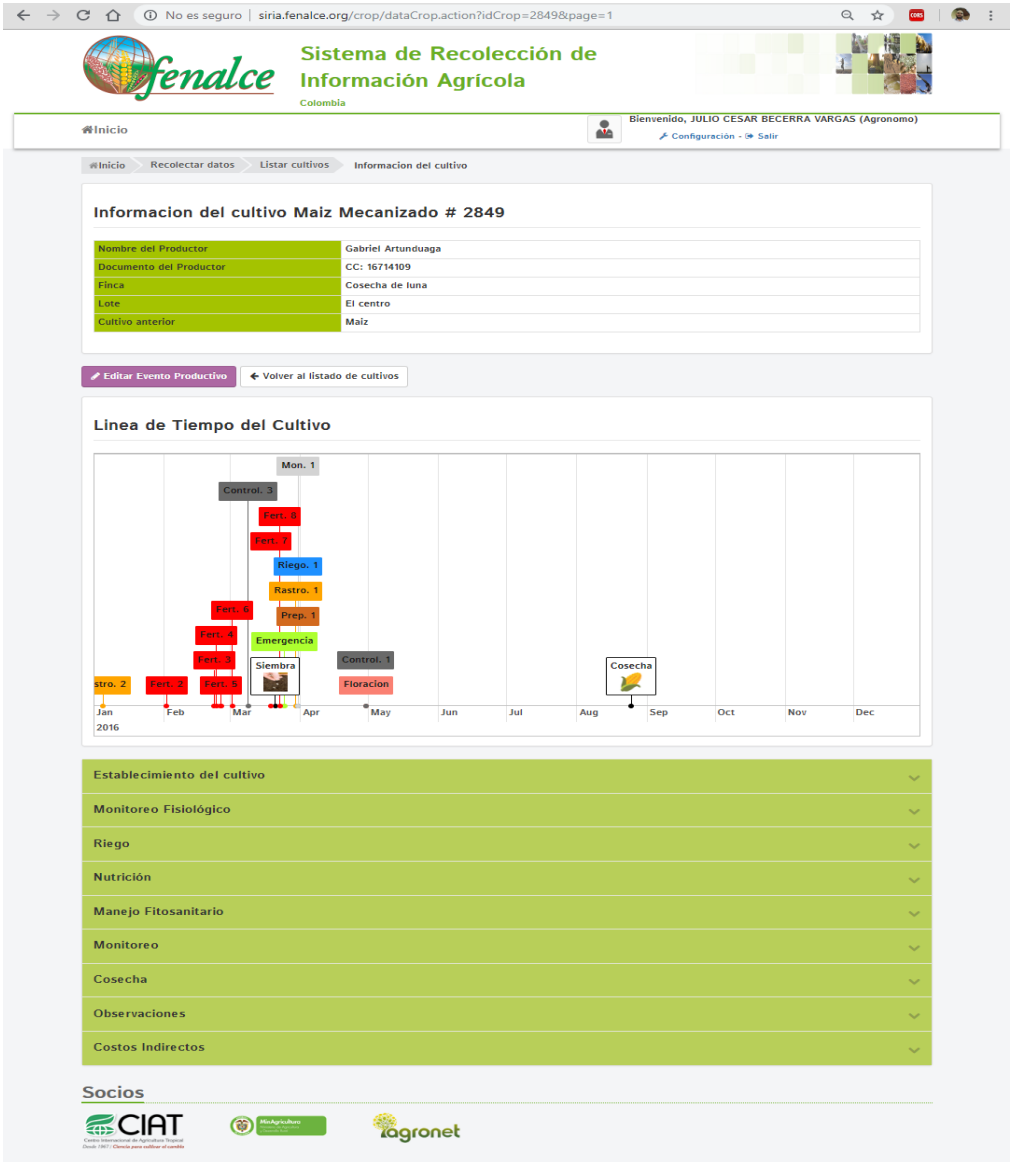
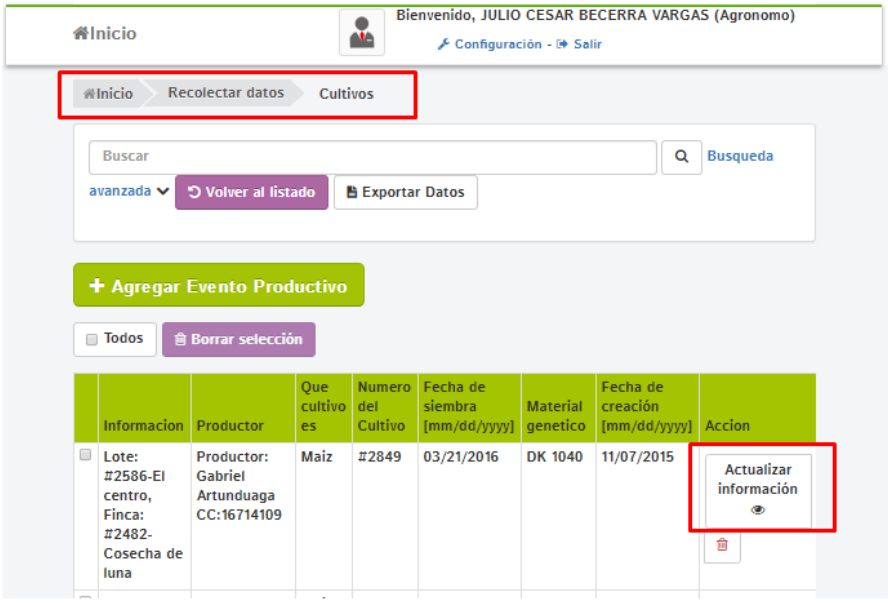


Figura 22. Formulario para registrar eventos de cultivo (SIRIA)

- Realizar bocetos del formulario web de la nueva aplicación como se observa en la **Figura 23** y **Figura 24**. Lo anterior partiendo de la familiarización con la interfaz de la plataforma de SIRIA y con el listado de variables que fueron seleccionadas en la vista minable, aquí se identificó que se conservaba, modificaba y adicionaba de los formularios de eventos de cultivos y caracterización de suelo encontrados en la plataforma.

http://siria.fenalce.org/
Entrar a:
Probar el sistema -> Recolectar datos -> Cultivos
Una vez aquí, escoger un cultivo de la lista y dar clic en la opción *actualizar información del cultivo*



Lo que se debe crear en la aplicación es el siguiente acordeón, pero solo aquellos ítems que NO estan subrayados

Figura 23. Captura 1 del boceto de formulario.

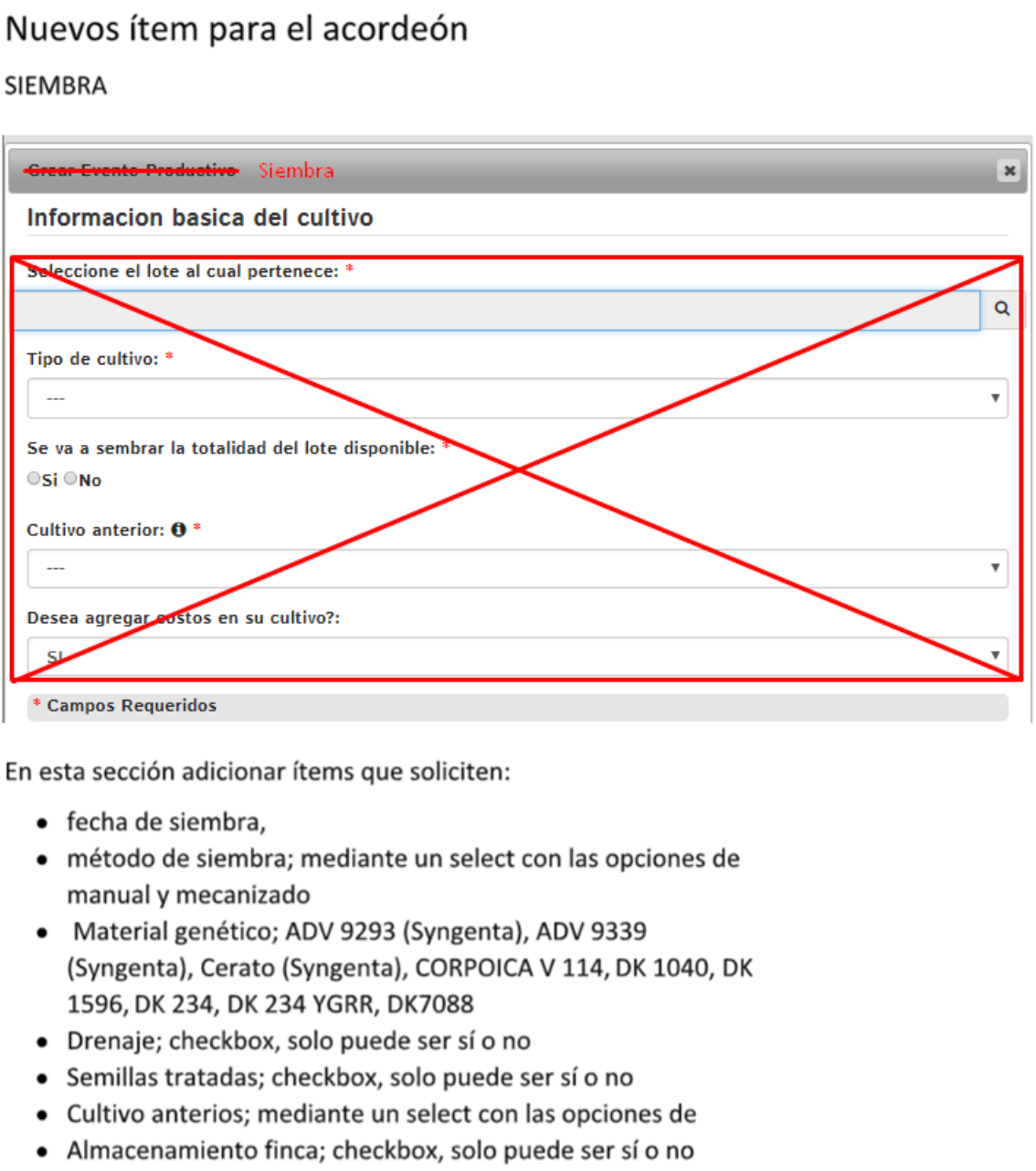


Figura 24. Captura 2 del boceto de formulario.

Establecer las tecnologías necesarias para desarrollar la aplicación web partiendo de que esta debería poder incorporarse o adaptarse por parte del cliente actual o futuros interesados (

Tabla 17).

Tabla 17. Tecnologías utilizadas en la aplicación web

	FRONT	BACK
LENGUAJE	HTML 5, CSS 3, JavaScript	Java
FRAMEWORK	Vue 2.0	Spring Boot
MOTOR BASE DE DATOS	Ninguno	Ninguno
COMUNICACIÓN	HTTP	HTTP

5.9.2.2 Sprint 2

En el segundo sprint se desarrolló el formulario web utilizando el framework Vue js y los bocetos definidos en el sprint 1 (**Figura 25**), los cuales conservaron aspectos visuales de la plataforma de SIRIA y contenían en su diseño entradas que permitían tratar las variables seleccionadas en la vista minable.

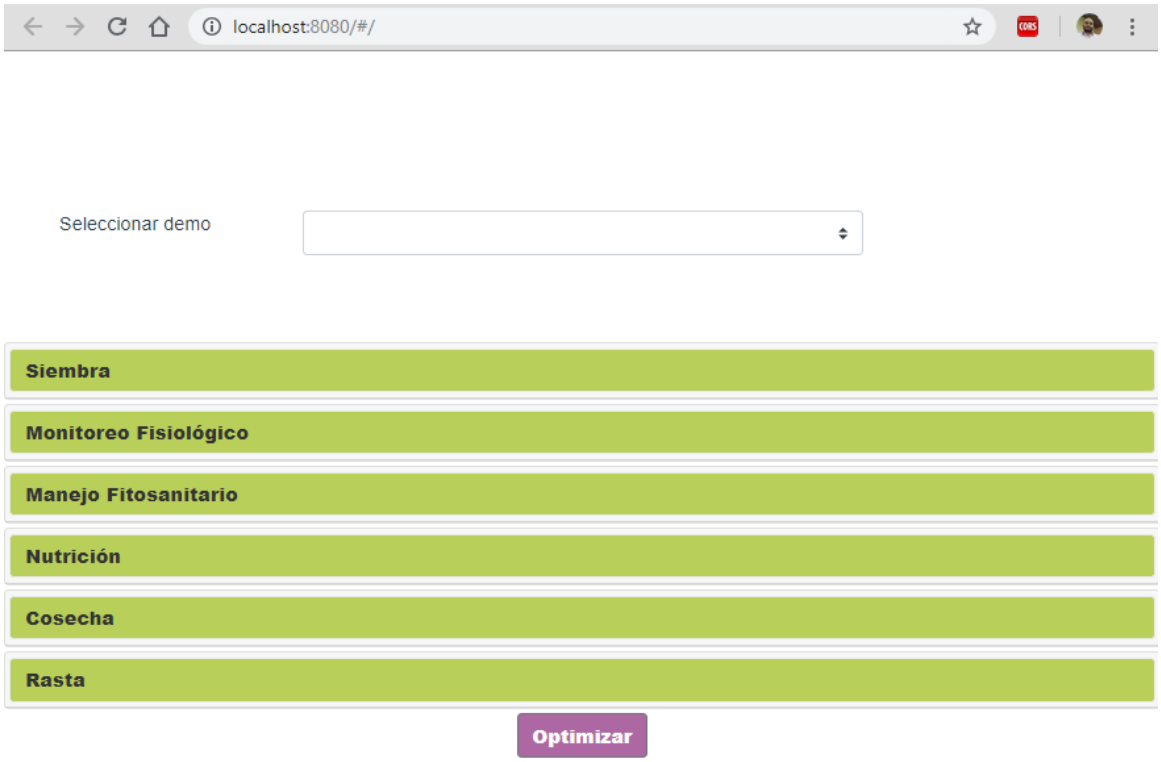


Figura 25. Formulario desarrollado

El formulario desarrollado conto con un listado de eventos de cultivos que cumple la función de ingresar valores en los campos del formulario y adicionalmente proporcionar información del clima a manera de ejemplo (**Figura 26**). Lo anterior teniendo en cuenta que al momento de la entrega de la aplicación web no se tendrá definido el servicio que obtendría los datos futuros del clima y que tampoco se espera contar con acceso a servicios que permitan recuperar información de caracterización de lotes.

El formulario se desarrolló para que una vez cargada la información del nuevo evento de cultivo se envié este a optimizar, las recomendaciones de optimización encontradas por GBHS se cargan remplazando la información suministrada por el usuario, adicionalmente en la parte superior del formulario se muestra el rendimiento que se espera obtenga el agricultor con las prácticas de manejo recomendadas y bajo las condiciones climáticas y del suelo a las que estará sometido el cultivo.

Finalmente se empaquetó la adaptación de la metaheurística GBHS desarrollada en java, esta se embebió junto con el servidor tomcat y los archivos html, css y javascript generados por Vue js, con lo que se generó una aplicación web auto contenida de tal forma que incluyera todo lo necesario para prestar su servicio evitando depender de un servidor de aplicaciones en el que desplegar la aplicación, con esto hace más fácil el despliegue en cualquier máquina.

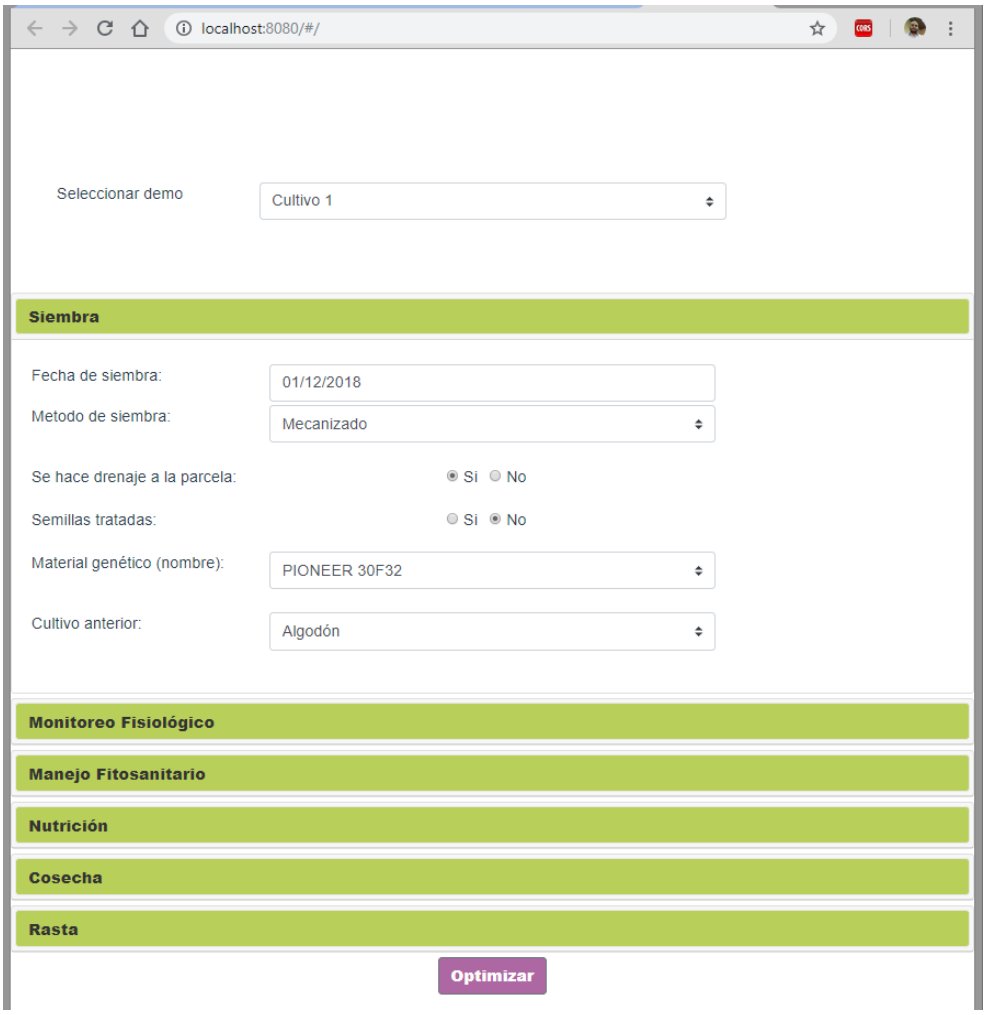


Figura 26. Formulario desarrollado con datos de ejemplo

5.10 FLUJO DE TRABAJO

A continuación se describe la manera en cómo se integran los capítulos 3, 4 y 5 para lograr realizar la caracterización de las zonas utilizadas en la siembra de cultivos de maíz al interior del departamento de Córdoba y con esto optimizar el rendimiento de estos.

Paso 1: Definir la vista minable apoyándose en lo realizado en el capítulo 3 donde se debe contar con información de eventos de cultivo de maíz realizados en el departamento de Córdoba como entrada.

Paso 2: Generar los clústeres con sus respectivos modelos utilizando alguna de las 2 adaptaciones propuestas (GRASP o MSSA) en el capítulo 4 y como entrada tomar la vista minable previamente definida para obtener agrupaciones de eventos de cultivos de maíz homogéneas a partir de sus características climáticas, de suelo y prácticas agronómicas desarrolladas.

Paso 3: Utilizar los clústeres y sus respectivos modelos que se generan en el paso anterior como base para el proceso de optimización realizado por GBHS definido en el capítulo 685 en donde se genera la clasificación y optimización de un evento de cultivo nuevo a ser desarrollado al interior del departamento del Córdoba.

Paso 4: Desplegar la aplicación web para que los usuarios puedan acceder al servicio de optimización (Paso 3), dentro de esta aplicación se encontrara un formulario en donde se cargara a partir de una base de datos la información asociada a la caracterización del terreno que será utilizado en el cultivos, con las coordenadas asociada al lote más la fechas indicada de inicio y fin se cargara información sobre el pronóstico del estado climático a presentarse durante la duración del cultivo. Con la información de clima y suelo cargada, el usuario debe decidir que campos asociados a prácticas de manejo como fertilización, controles, siembra, cosecha entre otras desea que se le sean optimizados y cuáles no, aquellos valores que se deseen optimizar se dejan sin diligenciar en el formulario y al final de la optimización se encontrara en el respectivo campo el valor que dentro de la optimización se consideró más relevante.

Nota: Los pasos fueron ejecutados en su respectivo orden para el presente trabajo de grado y en el momento de necesitar actualizar los resultados obtenidos debido a contar con un nuevo conjunto de datos se deben repetir nuevamente estos pasos con la nueva información disponible, la cual será la entrada del paso 1.

El personal encargado de realizar el despliegue debe integrar las bases de datos para recuperar la información de los terrenos e integrar los servicios de predicción climática.

CAPÍTULO 6

6 CONCLUSIONES Y TRABAJO FUTURO

6.1 CONCLUSIONES

6.1.1 Adaptación del enfoque clusterwise a cultivos de maíz

La adaptación del enfoque clusterwise para caracterizar las zonas utilizadas en la siembra de maíz al interior del departamento de Córdoba fue posible realizarla mediante el uso del algoritmo de k-means junto con la distancia euclidiana mixta para soportar la etapa de barajado inicial que define el enfoque, adicionalmente el uso de metaheurísticas como MSSA o GRASP en conjunto con la librería de Weka 3.8 para la generación de modelos de regresión lineal múltiple permite soportar el proceso de distribución de observaciones entre los K grupos en la etapa de optimización del enfoque. Al proporcionar como entrada a la adaptación del enfoque un conjunto de datos (vista minable) que contiene registros sobre eventos de cultivo de maíz realizados en el departamento de Córdoba se obtiene como salida K grupos de los que se generan las respectivas caracterizaciones gracias a los eventos pertenecientes a cada uno de estos.

Las soluciones propuestas con GRASP y MSSA lograron encontrar la mayor medida de calidad al agrupar las observaciones en 5 clústeres. No fue posible probar con un número mayor de 5 agrupaciones debido a que bajo los supuestos del modelo lineal clásico, el número de observaciones debe ser igual o mayor que el número de variables. A pesar de la limitante por las 115 variables y 799 observaciones recolectadas, la medida de calidad de los modelos encontrados por las adaptaciones propuestas del enfoque son mayor en un 13% frente a la encontrada si se procesan todas las observaciones de la vista minable de Córdoba como si pertenecieran a un único grupo.

Las soluciones propuestas con GRASP y MSSA presentan buenas medidas de calidad debido a que en promedio con 5 clústeres el R^2 ajustado toma el valor de 0.86, al someter las medias de las medidas de calidad de los resultados obtenidos por estas adaptaciones a la prueba estadística no paramétrica de Wilcoxon y a la prueba *t* para muestras apareadas se encontró que no existía diferencia estadísticamente significativa entre estas, esto debido a que las pruebas arrojaron un valor p de 0.08 y 0.23 respectivamente.

El macro modelo generado por la propuesta con GRASP y con la semilla 41 fue seleccionado por su buen R^2 ajustado promedio (0.88) y la simplicidad de los modelos asociados (56 variables por modelo en promedio), se observó que los modelos y las agrupaciones formadas permiten identificar que las características del suelo, el clima presentado y el manejo dado al cultivo, influyen de manera diferente en cada uno de estos grupos, lo que coincide con los principios de la agricultura especifica por sitio. Con esto se confirma que las soluciones propuestas están aportando a la identificación de prácticas menos genéricas, adicionalmente con los grupos obtenidos es posible realizar una caracterización de estos,

permitiendo crear planes de fertilización especifica por sitio sobre cultivos de maíz a realizarse al interior del departamento de Córdoba.

6.1.2 Optimización

La adaptación realizada a GBHS presentó un desempeño superior frente a la adaptación realizada a PSO, este comportamiento superior se presentó en todos los aspectos analizados, los cuales fueron:

- Promedios de porcentajes de optimización donde GBHS obtuvo en promedio 171.93% frente a 84.6% de PSO.
- Mayores valores de rendimientos encontrados en donde GBHS encontró combinaciones que generaba hasta 20 ton/ha y en su lugar PSO solo hasta 15 ton/ha.
- Menores tiempos de ejecución en donde PSO en promedio tarda 5 veces más que GBHS.
- Mayor cantidad de observaciones optimizadas en las pruebas en donde GBHS solo en una observación le fue imposible encontrar optimización alguna mientras que PSO presento este inconveniente en 17 observaciones.

Los valores de rendimiento encontrados por GBHS están dentro de lo posible, esta metaheurística encontró rendimientos máximos de 20 ton/ha y se conoce de cultivos realizados en México donde se ha reportado estos niveles de rendimiento [44].

6.2 TRABAJOS FUTUROS

En relación con la definición de la vista minable: Las mayores diferencias encontradas entre los clústeres fueron identificadas en la textura del suelo y la resistencia al rompimiento. En el proceso donde se realizó la división de la variable que contenía esta información dentro de varias columnas con valores porcentuales, no se conservó el orden de ocurrencia (desde la más superficial a la más profunda), este es un factor importante para esta característica debido a que indicaría a cuál de estas características tendría contacto la planta gracias a la profundidad de sus raíces, por lo anterior es de utilidad que en próximos trabajos se plantee una manera de conservar el orden de estas.

También, es importante que el CIAT y FENALCE continúen afianzando la cultura de buen manejo de la información por parte de los agricultores, con esto será posible aumentar el volumen de datos y la calidad de estos, con lo que se permitirá explorar las predicciones con otras técnicas como redes neuronales que no parten del supuesto de distribuciones conocidas ni de la linealidad de la relación entre los datos.

En relación con el proceso de clusterwise: Sería conveniente realizar nuevas investigaciones que involucren nuevas metaheurísticas en el proceso de búsqueda de la distribución óptima de *K* grupos de los eventos de cultivo. También se hace interesante definir nuevas estrategias para identificar el vecindario para un agrupamiento de observaciones, que en lo posible incluya mayor conocimiento del problema.

En relación con el proceso de optimización: Es necesario involucrar el manejo de costos (restricciones financieras), debido a que existen recomendaciones que generan buenos niveles de producción, pero por limitantes económicas no son viables de aplicar por parte de los agricultores, también es importante involucrar el impacto ambiental de uso de algunos fertilizantes (restricciones en fertilizantes).

BIBLIOGRAFIA

[1] CONPES, “Política para el Desarrollo Integral de la Orinoquia: Altillanura - fase I,” 2014. [Online]. Available: https://goo.gl/Zu8CGk. [Accessed: 15-Sep-2017].

[2] FENALCE, “Perspectivas de Maíz y Soya a 2022,” 2014. [Online]. Available: https://goo.gl/Jt7qDU. [Accessed: 15-Sep-2017].

[3] FENALCE, “Indicadores Cerealistas,” 2016. [Online]. Available: https://goo.gl/F4hTdc. [Accessed: 15-Sep-2017].

[4] FENALCE, “Productos que Representamos.” [Online]. Available: https://goo.gl/K5aSTc. [Accessed: 13-Sep-2017].

[5] FENALCE, “Área, Producción y Rendimiento Cereales y Leguminosas 2016 A,” 2016. [Online]. Available: https://goo.gl/JVQjrQ. [Accessed: 24-Oct-2017].

[6] FENALCE, “Área, Producción y Rendimiento Cereales y Leguminosas 2016 B,” 2016. [Online]. Available: https://goo.gl/ik4DkP. [Accessed: 24-Oct-2017].

[7] Gobernación de Antioquia - Secretaría de Agricultura y Desarrollo Rural, “Manual Técnico del Cultivo de Maíz Bajo Buenas Prácticas Agrícolas,” 2012. [Online]. Available: https://goo.gl/dFSKzU.

[8] FENALCE, “Producción de Harinas Precocidas de Maíz Plan de Negocios,” 2007. [Online]. Available: https://goo.gl/zh4i7T. [Accessed: 24-Oct-2017].

[9] Superintendencia de Industria y Comercio, “Cadena Productiva del Maíz.” [Online]. Available: https://goo.gl/dYbRms.

[10] J. Rodríguez, A. M. González, F. Rodrigo, and L. Guerrero, “Fertilización por Sitio Específico en un Cultivo de Maíz (Zea mays L .) en la Sabana de Bogotá,” *Agronomía Colombiana*, vol. 26, no. 2, Bogota, pp. 308–321, 2008.

[11] DANE, “Encuesta Experimental Nacional de Desempeño Agropecuario,” 2008. [Online]. Available: https://goo.gl/iMr2Ts. [Accessed: 15-Sep-2017].

[12] DANE, “Cuenta Satélite Piloto de la Agroindustria (CSPA): Maíz, Sorgo y Soya y su Primer Nivel de Transformación,” 2005. [Online]. Available: https://goo.gl/NjXwvC. [Accessed: 08-Nov-2017].

[13] J. Cock, “Site Specific Agriculture Based on Sharing Farmers Experiences,” Cali, 2007.

[14] D. Jiménez *et al.*, “Interpretation of Commercial Production Information: A Case Study of Lulo (Solanum quitoense), an Under-Researched Andean Fruit,” *Agric. Syst.*, vol. 104, no. 3, pp. 258–270, Mar. 2011.

[15] D. Jiménez *et al.*, “From Observation to Information: Data-Driven Understanding of on Farm Yield Variation,” *PLoS One*, vol. 11, no. 3, p. e0150015, Mar. 2016.

[16] D. Jiménez *et al.*, “Analysis of Andean Blackberry (Rubus glaucus) Production Models Obtained by Means of Artificial Neural Networks Exploiting Information Collected by Small-Scale Growers in Colombia and Publicly Available Meteorological Data,” *Comput. Electron. Agric.*, vol. 69, no. 2, pp. 198–208, Dec. 2009.

[17] H. Dorado, “Analisis Multivariado para la Caracterizacion de Factores Edafo Climaticos y Agrupamientos de Sitios con Presencia de Platano,” Universidad del Valle, 2012.

[18] V. Criollo B., “Respuesta del Cultivo de Maíz (Zea mays L.) a la Fertilización por Sitio Específico en Suelos De La Sabana de Bogotá,” Universidad Nacional de Colombia, 2009.

[19] E. Ismail, M. Hussien, R. Hamed, and A. El-Hefnawy, “A Mathematical Programming Approach to Variable Selection in Clusterwise Linear Regression,” Cairo University, 2012.

[20] M. Khadka and A. Paz, “Comprehensive Clusterwise Linear Regression for Pavement Management Systems,” *J. Transp. Eng. Part B Pavements*, vol. 143, no. 4, p. 04017014, Dec. 2017.

[21] M. S and M. E, “An Analysis on Clustering Algorithms in Data Mining,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 1, pp. 334–340, 2014.

[22] H. Spiith, “Algorithm 39 Clusterwise linear regression,” *Computing*, vol. 22, no. 4, pp. 367–373, 1979.

[23] H. Späth, “Correction to: ‘Algorithm 39: clusterwise linear regression’” [Computing (1979), no. 4, 367-373],’” *Computing*, vol. 26, no. 3, p. 275, 1981.

[24] J. Gorgas García, N. Cardiel López, and J. Zamorano Calvo, *Estadística Básica para Estudiantes de Ciencias*. Departamento de Astrofísica y Ciencias de la Atmósfera, Universidad Complutense de Madrid, 2009.

[25] D. M. Levine, T. C. Krehbiel, M. L. Berenson, M. L. González Díaz, and S. A. Durán Reyes, *Estadística para Administración*. Pearson/Educación, 2006.

[26] A. D. Martínez and C. R. Portuondo Padrón, “System Modeling in Organizations Using Artificial Intelligence. Application in Management,” *Retos la Dir.*, vol. 10, no. 1, pp. 1–18, 2016.

[27] M. C. Vélez and J. A. Montoya, “Metaheurísticos: Una Alternativa Para la Solución de Problemas Combinatorios en Administración de Operaciones,”

Rev. Esc. Ing. Antioquia, pp. 99–115, 2007.

[28] Z. Beheshti and S. M. H. Shamsuddin, “A Review of Population-Based Meta-Heuristic Algorithm,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 5, no. 1, pp. 1–35, 2013.

[29] J. E. Rowe, “Genetic Algorithms,” in *Springer Handbook of Computational Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 825–844.

[30] R. Marti and M. Moreno Vega, “MultiStart Methods,” *Rev. Iberoam. Intel. Artif.*, vol. 19, pp. 49–60, 2003.

[31] M. G. C. Resende and C. C. Ribeiro, “GRASP: Greedy Randomized Adaptive Search Procedures,” in *Search Methodologies*, Boston, MA: Springer US, 2014, pp. 287–312.

[32] V. G. Chamorro and B. Barán, “Optimización por Enjambre de Partículas para Satisfacción de Fórmulas Booleanas,” *34th Latin-American Conf. Informatics*, 2010.

[33] E. Cuevas and N. Ortega-Sánchez, “El Algoritmo de Búsqueda Armónica y sus Usos en el Procesamiento Digital de Imágenes,” *Comput. y Sist.*, vol. 17, no. 4, pp. 543–560, Dec. 2013.

[34] C. Cobos, J. Perez, and D. Estupiñan, “A Survey of Harmony Search.Una Revisión de la Búsqueda Armónica.,” *Rev. Av. en Sist. e Informática (RASI).*, vol. 8, no. 2, pp. 67–80, Jul. 2011.

[35] J. Yadav and M. Sharma, “A Review of K-mean Algorithm,” *Int. J. Eng. Trends Technol.*, vol. 4, no. 7, pp. 2972–2976, 2013.

[36] Y. Miao, D. J. Mulla, and P. C. Robert, “Identifying Important Factors Influencing Corn Yield and Grain Quality Variability Using Artificial Neural Networks,” *Precis. Agric.*, vol. 7, no. 2, pp. 117–135, May 2006.

[37] P. TITTONELL, K. SHEPHERD, B. VANLAUWE, and K. GILLER, “Unravelling the Effects of Soil and Crop Management on Maize Productivity in Smallholder Agricultural Systems of Western Kenya—An Application of Classification and Regression Tree Analysis,” *Agric. Ecosyst. Environ.*, vol. 123, no. 1–3, pp. 137–150, Jan. 2008.

[38] A. Shekoofa, Y. Emam, N. Shekoufa, M. Ebrahimi, and E. Ebrahimie, “Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture,” *PLoS One*, vol. 9, no. 5, p. e97288, May 2014.

[39] S. Delerce *et al.*, “Assessing Weather-Yield Relationships in Rice at Local Scale Using Data Mining Approaches,” *PLoS One*, vol. 11, no. 8, p. e0161620, Aug. 2016.

[40] O. E. Quinteros, A. Funes, and H. C. Ahumada, “Construcción de una vista minable para aplicar minería de datos secuenciales temporales,” 2016.

[41] G. Biau, “Analysis of a Random Forests Model,” *Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2010.

[42] F. J. Fabozzi, S. M. Focardi, S. T. Rachev, and M. Selection, “Model Selection Criterion : AIC and BIC,” vol. 41, no. 1979, pp. 399–403, 2014.

[43] N. M. Betzek, E. G. de Souza, C. L. Bazzi, K. Schenatto, and A. Gavioli, “Rectification Methods For Optimization of Management Zones,” *Comput. Electron. Agric.*, vol. 146, pp. 1–11, Mar. 2018.

[44] AgroSíntesis, “Cómo Producir 20 ton/ha de Maíz,” 2016. [Online]. Available: https://goo.gl/t3JbJu.