



## **GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM**

**CÓDIGO: GIN-GUI-001**

**VERSIÓN: 01**

**FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015**



**UNIVERSIDAD AUTÓNOMA DE MANIZALES**

**VICERRECTORÍA ACADÉMICA**

**UNIDAD DE INVESTIGACIÓN**

**UNIDAD DE POSGRADOS**



## GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM

CÓDIGO: GIN-GUI-001

VERSIÓN: 01

FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015

### TÓPICOS PARA LA PRESENTACIÓN DE INFORMES FINALES<sup>1</sup>

UNIVERSIDAD AUTÓNOMA DE MANIZALES

PROYECTO: CLASIFICACIÓN DE ELEMENTOS TRANSPONIBLES MEDIANTE  
TÉCNICAS DE MACHINE LEARNING Y AFINAMIENTO DE HIPERPARÁMETROS

GRUPO DE INVESTIGACIÓN: INGENIERÍA DEL SOFTWARE Y AUTOMÁTICA

ESTUDIANTE: JOHAN SEBASTIAN PIÑA DURAN

TUTOR DE TESIS: SIMÓN OROZCO ARIAS


CO-TUTOR DE TESIS: REINEL TABARES SOTO

DATOS DE IDENTIFICACIÓN: 1.001.851.993

AÑO: 2020

---

<sup>1</sup>

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


## 1. RESUMEN

El presente proyecto de investigación está enfocado en la creación de una estrategia para la clasificación de elementos transponibles a nivel de linajes mediante algoritmos de aprendizaje automático y usando técnicas de afinamiento de hiperparámetros y ensamble de algoritmos. El propósito de este proyecto es entender a profundidad los genomas de interés agrícola en Colombia como el arroz, el café, la piña o la caña de azúcar, y posiblemente ser aplicado a otros genomas.

Para el desarrollo de este proyecto se utilizan los algoritmos comunes de aprendizaje supervisado, así como métodos de pre-procesamiento y métricas para evaluar el desempeño de la clasificación. El conjunto de datos con los que se desarrolló el proyecto fueron provistos por el tutor. Cada algoritmo de aprendizaje supervisado primero fue entrenado con los parámetros por defecto, posteriormente, se establecieron los valores con los cuales se iban a iterar cada uno de los hiperparámetros para encontrar de esta forma la combinación de estos que proporcionara el mejor de los desempeños. Finalmente, con los algoritmos afinados se aplicó un algoritmo de ensamble con el objetivo de lograr desempeños más altos que los obtenidos con los algoritmos entrenados independientemente.

Con el tuneo de los algoritmos se lograron obtener desempeños superiores al 95% (en F1-Score) de efectividad utilizando como esquemas de pre-procesamiento en el conjunto de datos reducción de dimensionalidad usando el análisis de componentes principales y escalamiento. Adicionalmente, se logró con los métodos de ensamble un desempeño de 97% en la clasificación de los elementos transponibles.

**PALABRAS CLAVES:** *Machine Learning*, Elementos transponibles, Bioinformática, Métricas, Clasificación, LTR retrotransposones.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## ABSTRACT

This research project is oriented to proposed a novel strategy for transposable elements classification at linages level using techniques of hyperparameters tuning and algorithm ensembles in order to have a deeper understanding of plants genomes in Colombia used for agriculture. For example, rice, coffee or pineapple and eventually be applied in others genomes.


To develop this project, common algorithms of Machine learning are used as same as pre-processing methods and metrics to evaluate the classification performance. The dataset used in this project is provided by research mentor. Each machine learning algorithm first was trained with default hyperparameters, then, ranges of these hyperparameters were defined and in order to found the combination of them that provide the best performance. Finally, an ensemble algorithm was trained using the tuned algorithms to achieve a higher performance for transposable elements classification than algorithms trained individually.

Tunning hyperparameters is a useful technique to enhance accuracies of machine learning models and specially for transposable elements classification these models achieve performances up to 95% (F1-Score) using dimensional reduction with principal component analysis and data scaling. In addition, ensemble methods show a performance of 90% for transposable elements classification.

**KEY WORDS:** Machine learning, Transposable elements, Bioinformatics, Metrics, Classification, LTR retrotransposons.

## TABLA DE CONTENIDO


1. RESUMEN .....	3
2. PRESENTACIÓN.....	5
3. INTRODUCCIÓN .....	6
4. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN .....	7
5. REFERENTE TEÓRICO .....	9

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

<b>6. LOS OBJETIVOS .....</b>	<b>13</b>
<b>7. METODOLOGÍA.....</b>	<b>13</b>
<b>8. RESULTADOS .....</b>	<b>16</b>
<b>9. DISCUSIÓN DE RESULTADOS .....</b>	<b>21</b>
<b>10. CONCLUSIONES.....</b>	<b>22</b>
<b>11. RECOMENDACIONES.....</b>	<b>22</b>
<b>12. EVIDENCIA DE RESULTADOS EN GENERACIÓN DE CONOCIMIENTO, FORTALECIMIENTO DE LA CAPACIDAD CIENTÍFICA Y APROPIACIÓN SOCIAL DEL CONOCIMIENTO, FORMACIÓN .....</b>	<b>23</b>
<b>13. IMPACTOS LOGRADOS .....</b>	<b>23</b>
<b>14. BIBLIOGRAFÍA.....</b>	<b>24</b>
<b>15. ANEXOS .....</b>	<b>33</b>

## 2. PRESENTACIÓN

Las nuevas tecnologías de secuenciación y la gran cantidad de datos que han sido liberados en los últimos años han permitido que las investigaciones en temas bioinformáticos tomen un enfoque diferente, centrando su atención principalmente en el análisis de esos grandes volúmenes de datos para la interpretación de la información contenida en ellos con el fin de entender como está dispuesta la información genética de las diversas especies a las que se tiene acceso mediante esta información. Dentro de las diferentes estructuras que actualmente se han podido identificar en el ADN (ácido desoxirribonucleico) se encuentran los elementos transponibles (TEs), elementos repetitivos los cuales llegan a ser hasta el 80% del material genético en las plantas y que son los responsables de las diversidades de las especies incluso entre las especies más cercanas; En Colombia, debido a la diversidad de especies que se tienen y el interés agronómico que estas contemplan, la identificación y


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

anotación de los elementos transponibles presentes en especies como café, piña, arroz, entre otros, permiten brindar información a la sociedad científica que pueda ser utilizada en el mejoramiento genético a través de técnicas de ingeniería genética o biotecnología logrando ventajas como mayor producción, menores pérdidas debido al cambio climático y mayor calidad en los productos finales. Adicionalmente, el uso de técnicas de machine Learning (ML) para el análisis de las bases de datos biológicas, especialmente las bases de datos de TEs, es novedoso y un campo que solo hasta hace algunos años se ha venido explorando, diferentes softwares basados en ML han sido diseñados para detectar elementos repetitivos, clasificarlos (a nivel de superfamilias), o ambos. Además, programas basados en redes neuronales profundas se han desarrollado para clasificar TEs. Sin embargo, no existen estudios que realicen la clasificación de estos TEs a nivel de linajes con ML y que además apliquen técnicas para ajustar estos modelos a la distribución de los datos, logrando eficiencias más altas con una métrica que establezca el grado de generalización de cada uno de los modelos. En este sentido, el proyecto busca realizar la clasificación de elementos transponibles a nivel de linajes presentes en diversas especies de plantas utilizando técnicas de inteligencia artificial, utilizando para tal fin, técnicas de afinamiento de hiperparámetros y métodos de ensamble con el fin de mejorar las eficiencias de los algoritmos aplicados.

### 3. INTRODUCCIÓN

Los elementos transponibles (TEs, pos sus siglas en inglés) son unidades genéticas que se encuentran en el ADN de los seres vivos y que tienen la habilidad de replicarse y multiplicarse en el genoma (1). Desde su descubrimiento, los elementos transponibles han mostrado una gran relevancia en la evolución de los genomas, modificando sus arquitecturas, diversidad y regulación (2–5). Los elementos transponibles son responsables de muchas de las variaciones en el fenotipo y el genotipo de las especies y pueden llegar a constituir en plantas la mayoría del genoma como en el de maíz en el que hasta el 85% de todo el material genético está identificado como material transponible (6). Los TEs se clasifican tradicionalmente en Clase I (retrotransposones) o Clase II (DNA transposones) (7). Adicionalmente, la clasificación de los elementos transponibles puede darse en otros niveles: (clases, subclases, ordenes, autonomía, superfamilias y linajes) (7–9).

Usualmente, existen varios métodos para la detección y clasificación de los elementos transponibles en genomas de organismos secuenciados, dentro de los más populares podemos encontrar: de *novo*,

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

basado en estructura, por genómica comparativa y basado en homología (10). Estos métodos ofrecen especificidad y sensibilidad, pero todos tienen una alta tasa de detección de falsos positivos (11). Otros métodos reportados se basan en el ensamblaje de las lecturas repetitivas como RED (12), TE<sub>deno</sub> (13), Transposome (14), and REPdenovo (15). Por otro lado, LTRClassifier (16) y Inpactor (17) solo realizan la clasificación de TEs detectados usando otros métodos.


El *Machine learning* (ML) es definido como el conjunto de métodos computacionales que usan la experiencia para optimizar y mejorar el desempeño de los criterios de decisiones para construir modelos que tengan la habilidad de detectar, clasificar o predecir nuevos datos basados en experiencias pasadas. El ML ha sido aplicado a numerosos problemas bioinformáticos como la genómica (18), biología de sistemas y evolución (19) demostrando grandes alcances en precisión y velocidad. Para el caso de la clasificación de TEs se reportan considerables mejoras de los resultados obtenidos (20,21). Sin embargo, estas clasificaciones se llevan a cabo en niveles poco específicos de los TEs y a nivel de linajes los resultados publicados no tienen una alta precisión.

Es por esto que el presente proyecto de investigación propone la creación de una estrategia para la clasificación de los elementos transponibles a nivel de linajes con la ayuda de algoritmos tradicionales de machine learning y técnicas de preprocesamiento, métricas y estrategias como el tuneo de parámetros y ensamble de algoritmos para lograr desempeños más altos en efectividad y generalización de los algoritmos.

#### 4. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN

Debido al cambio climático, la economía de países productores se ve afectada en gran medida por la pérdida de los cultivos (22). Millones de personas en el mundo tienen como principal actividad económica la producción agrícola y debido a los efectos de esta problemática han tenido que cambiar de actividad económica. Para mitigar estos efectos, se requiere el desarrollo de variedades o especies resistentes a condiciones agrestes del cambio climático (23). Para lograr esto se requiere un amplio entendimiento de las especies que son ampliamente cultivadas, especies que en Colombia son también de gran interés agrícola como el café, la caña de azúcar, la papa, el maíz, entre otros (24). Este conocimiento asegura mayor productividad en el sector agrícola, con cosechas más




	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

resistentes a plagas, menores tiempos entre la siembra y cosecha y adaptación al medioambiente.(11)

A partir de la develación del ADN (Ácido desoxirribonucleico) y su importancia en el almacenamiento y replicación del material genético se ha descubierto que la mayoría de las funciones de las células están contenidas en el ADN y son transmitidas de generación en generación. Esta información ha causado una gran revolución en el entendimiento de la biología de las especies y ha sido de gran utilidad a la hora de crear variedades mejor adaptadas a los cambios climáticos (25). Uno de los hechos claves que ha impulsado esta revolución fue el desarrollo y posterior mejora de las tecnologías de secuenciación, que han producido gran cantidad de secuencias de ADN de muchas especies de interés para su posterior estudio *in silico* (26) A través de técnicas de bioinformática, se ha logrado procesar, analizar y posteriormente entender muchas de las funciones que son vitales en los organismos, a través de un pipeline que se ha vuelto estándar en los últimos años. El primer paso en este pipeline es la secuenciación, en donde a través de diversas metodologías se obtiene una muestra biológica de ADN o ARN, se cortan en múltiples porciones y se duplican para mejorar la calidad, obteniendo un archivo informático con la información genómica contenida en la muestra. A continuación, se lleva a cabo el ensamblaje de las lecturas que son obtenidas en el paso anterior; este proceso intenta agrupar y formar de nuevo las estructuras cromosómicas originales. Por último, se anotan todas las características de interés como los genes y sus funciones, porciones codificantes y no codificantes, secuencias promotoras, elementos transponibles, entre otros. Aunque en la actualidad muchas especies de plantas ya han sido totalmente secuenciadas y estudiadas a nivel molecular, muchos de los cultivos de interés como el maíz y la caña de azúcar, aún conservan muchas secciones de sus genomas que son desconocidas. Estos vacíos son generados principalmente por secuencias repetitivas como los TEs y las repeticiones simples como micro satélites, debido a que la mayoría de los ensambladores tienen problemas con regiones que son muy repetitivas y muy variables (27). Además, los TEs pueden afectar la anotación de los genes (28), por lo que se recomienda identificarlos y ocultarlos antes de ejecutar el proceso de anotación.

Investigaciones han demostrado que los TEs tienen funciones claves en muchos aspectos como en el aporte a la diversidad intra-especie (26), influencia sobre los genes (27,29) y en la adaptación a cambios climáticos (30). Por esto, ha surgido la necesidad de entender a profundidad las dinámicas



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


de estas secuencias, pero su clasificación es un reto para la genómica y la bioinformática actual ya que, debido a su naturaleza repetitiva, sus abundantes polimorfismos, su especificidad en cada especie y su gran divergencia inclusive entre especies de plantas cercanas, hacen que los resultados sean poco fiables (31). existen diversos algoritmos que implementan técnicas bioinformáticas y de analíticas de datos (14,16,17,32) pero sus tasas de falsos positivos son muy altas (33). Por lo tanto, se requieren estrategias novedosas para mejorar el proceso de descubrimiento de TEs y así mejorar el entendimiento a niveles moleculares de plantas e inclusive otros organismos.

El ML ofrece la ventaja de optimizar tareas usando experiencia previa. Gracias a la gran cantidad de datos que se han generado en los últimos años, muchos investigadores han aplicado técnicas de ML en diversas áreas de la genómica, particularmente en la identificación y clasificación de TEs. Por ejemplo, (21) probaron diversos algoritmos de ML tales como redes neuronales, redes bayesianas, *Random Forest* y árboles de decisión para mejorar los resultados de identificación o clasificación usando como entrada las salidas de varias herramientas bioinformáticas. Aunque este trabajo mostró muy buenos resultados, los investigadores usaron datos simulados, lo que puede ser impráctico si se lleva a una aplicación real. (34) desarrollaron TEclass para clasificar elementos transponibles a través de máquinas de soporte vectorial (SVM por sus siglas en inglés), pero solo hasta el nivel de órdenes.

Se requiere por tanto el diseño, implementación y validación usando datos de TEs reales de una arquitectura basada en ML, que mejore la identificación y clasificación de los TEs a niveles de súper familias y linajes, con el objetivo de mejorar los conocimientos sobre su diversidad, sus dinámicas y sus impactos en los genomas de plantas. Lo anterior con el fin de sentar las bases para la posterior mejora de variedades de cultivos de interés agronómico mundial, tales como el arroz, el maíz, la caña de azúcar, el trigo, la cebada o el café, aportando soluciones a las pérdidas de cosechas debido al cambio climático.

## 5. REFERENTE TEÓRICO

Los elementos transponibles (TEs) fueron descubiertos por primera vez por Barbara McClintock al experimentar con plantas de maíz en 1944 (35). Actualmente, se sabe que estos cumplen un papel importante en los genomas de muchas especies y que pueden componer una gran parte de la información genética. (36). Los retrotransposones son un tipo de elemento transponible de los más


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

abundantes en las especies debido a que tienen un mecanismo de copia y pega usando el ARN lo que aumenta significativamente la presencia de estos elementos en el material genético (37,38). Por ejemplo, en el maíz estos elementos pueden componer hasta el 80% de todo el mapa genético (39). Esto es especialmente cierto en el caso de los LTR retrotransposones.

Los TEs ellos tienen diferencias significativas en cuanto a su estructura, la presencia de regiones reguladoras enzimáticas y en su ciclo de vida (40). En un principio, a los TEs se les atribuían funciones negativas (34), pero diferentes investigaciones han demostrado tener un papel clave en muchos procesos genéticos (41), como: reorganización del genoma (42), promoción de la expresión génica masculina en la espermatogénesis tardía (43), organización de los cromosomas, en particular en los cromosomas sexuales, participación en eventos de reordenamiento (44,45) (p. ej., translocaciones, fusiones o fisuras), y contribución a las variaciones de tamaño del genoma (29).

Debido a su alta presencia en los genomas, caracterizar y clasificar los elementos transponibles es fundamental para comprender la dinámica y los mecanismos de la evolución del genoma (29,46–48). Además, la anotación de los TEs puede mejorar la precisión de la anotación de las regiones codificantes y facilitar los estudios genéticos funcionales (49), basándose en el desarrollo de diferentes estrategias de identificación y clasificación bioinformática automática.


En los últimos años se ha realizado un esfuerzo considerable para crear un sistema unificado de clasificación. Uno de los métodos más aceptados fue el sistema de clasificación jerárquica que subdividió a los TEs en clases, subclases, órdenes, súper-familias, linajes y familias propuesto por Wicker y Keller en 2007 (49,50). Siguiendo la propuesta de Wicker, los TEs pueden clasificarse en clases según su mecanismo de replicación, y pueden dividirse en retrotransposones (Clase 1) y transposones de ADN (Clase 2) (51). En el ADN, hay elementos que son rodeados por medio de repeticiones terminales largas (LTR) y se denominan LTR retrotransposones. Otros retrotransposones carecen de LTR (no-LTR retrotransposones) y están separados en dos grupos, elementos intercalados largos (LINEs) y elementos intercalados cortos (SINEs) (51). Los LTR-RTs de plantas pueden agruparse en dos súper-familias, Copia y Gypsy, que difieren por la posición en el dominio INT en la poliproteína (48,52). También es posible dividir estas súper-familias en linajes.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

Dado que los TEs están bajo una presión de selección relativamente baja y evolucionan más rápidamente que los genes codificadores (53), su clasificación y anotación es una tarea muy compleja (33). Se han hecho muchos intentos para crear un sistema unificado de clasificación que combine los aspectos filogenéticos y enzimáticos, sin embargo, la clasificación se vuelve más difícil en los niveles más bajos como las súper-familias y los linajes (39). En algunos casos, se requiere de una investigación compleja y manual hecha por especialistas. No existe una sola herramienta que pueda aplicarse universalmente a todas las especies para todos los tipos de TEs (54), por lo tanto, en la literatura se pueden encontrar muchas técnicas, métodos y software diferentes. De esta manera, existen diversas formas de agrupar técnicas o métodos para identificar los TEs. La mayoría de los autores han propuesto algunas de las siguientes categorías (7,53,55,56): basada en estructuras, basada en homología, de *novο* y genómica comparativa. Además, muchas herramientas aplican más de un método para mejorar sus resultados (57).

Los algoritmos que detectan la presencia de TEs basados en estructura buscan las características estructurales (21,58). Los métodos basados en homología detectan los TEs sobre la base de su similitud con las secuencias de TE de referencia (41,59). Cuando se dispone de una librería de TEs o de una base de datos de repetición para las especies estudiadas, el proceso de identificación puede ser sencillo. La creación de una librería para este método puede adquirirse de dos maneras: a través de bases de datos existentes o librerías construidas por métodos de *novο* u otros métodos (60). Esto se puede lograr utilizando cualquier herramienta de alineación de secuencias, como BLAST, que encontrará TEs con un valor de similitud superior a un umbral (57), o con RepeatMasker (61). A nivel linaje y familias, los enfoques filogenéticos (basados en la homología) son los más utilizados (49).

Uno de los enfoques de identificación de *novο* busca secuencias similares encontradas en muchas posiciones dentro de una secuencia (56) aprovechando la naturaleza repetitiva de los TEs (12). Esta forma es a través del conteo exacto o aproximado (conocido como "espaciado") de k-mers (12,58). Este método se denomina de *novο*, ya que no requiere información adicional sobre las secuencias de consulta (55). Sin embargo, los TEs con un bajo número de copias pueden no ser reconocidas como secuencias repetidas en este enfoque. Los consensos estructurales de los elementos de cada grupo de LTR retrotransposones son similares a los retrovirus excepto por la ausencia del gen *env* en muchos de los elementos. Todos los LTR retrotransposones contienen genes *gag* y *pol* que se


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

sobreponen en diferentes lecturas, o pueden estar separados por uno o más codones. Hay muchos ejemplos en todos los linajes, sin embargo, estos tienen el gen *gag* y *pol* se unen en un solo marco abierto de lectura (ORF, por sus siglas en inglés) (62).

Muchas aplicaciones bioinformáticas se han desarrollado siguiendo las estrategias antes mencionadas y la mayoría de ellos sólo pueden identificar clases específicas (retrotransposones o transposones de ADN) u órdenes como LTR-RTs o no-LTR retrotransposones. Aunque la minería de datos (63) y las técnicas de aprendizaje automático han mostrado resultados muy exitosos en otras tareas genómicas, muy pocas herramientas para los TEs aplican estas técnicas computacionales en sus algoritmos.

El aprendizaje automático (ML) puede definirse como el conjunto de técnicas que permiten a los algoritmos aprender basados en experiencia pasada para predecir nuevos datos por medio de la optimización de funciones de costo (64). El objetivo principal de las tareas de ML es optimizar una función de costo que se encuentra en términos de algunos parámetros de un modelo. En el proceso de optimización se calibra el modelo propuesto. Con este objetivo, los datos dados se dividen aleatoriamente en al menos tres subconjuntos: conjuntos de entrenamiento, validación y prueba (65). Este proceso es crucial para evitar el sobre ajuste o el sub ajuste. Por lo tanto, el algoritmo debe contener un balance entre la flexibilidad del modelo y la precisión del mismo ya que modelos muy flexibles pueden incurrir en una poca generalización de los datos y un modelo con alto porcentaje de efectividad y poco flexible puede generar un modelo que no es capaz de generalizar su clasificación para nuevas instancias (65).

El diseño e implementación de un sistema de ML es un proceso complejo que puede realizarse en tres pasos: 1) pre-procesamiento de datos brutos (por ejemplo, selección y extracción de características, imputación de datos, etc.), 2) aprendizaje o formación del modelo mediante el uso de un algoritmo o arquitectura ML apropiado (calibrar el modelo) y 3) evaluación del modelo mediante métricas (66). En algunos casos, el pre-procesamiento se debe hacer de forma semiautomática y guiada por expertos. Las técnicas de ML han obtenido mejores resultados que los métodos tradicionales en el problema de la clasificación de los TEs. A través de ML es posible clasificar LTR retrotransposones utilizando otras características además de las regiones de codificación (que se utilizan comúnmente en los procesos de clasificación), como la longitud del elemento, la longitud de

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

los LTR y las longitudes de los ORF (67). Los algoritmos ML son capaces de distinguir entre LTRs retrovirales y SINEs, combinando un conjunto de métodos (68), que es un procedimiento complejo en bioinformática. Además, utilizando la clasificación jerárquica, los métodos basados en ML demostraron obtener mejores resultados que los métodos basados en homólogos bien conocidos (específicamente el algoritmo de BLASTN) (69), sin embargo, la clasificación solo fue llevada hasta el nivel de súper-familias y no se abordó el problema de la identificación lo que resta especificidad en la clasificación.

Debido a que los TEs están bajo una presión de selección relativamente baja y evolucionan más rápidamente que los genes codificadores (53), están sujetos a una evolución acelerada debido a las recombinaciones y a la presencia de TEs anidados que dan lugar a nuevos elementos y variaciones de TEs (70), lo que hace que su clasificación y posterior anotación sea una tarea muy compleja (33). Por lo tanto, los métodos actuales en la bioinformática no pueden obtener resultados fiables en las tareas de detección y clasificación de TEs y tampoco han sido fiables en una clasificación a niveles de ordenamiento más específicos como a nivel de linajes o superfamilias.

## 6. LOS OBJETIVOS

El presente proyecto tiene como objetivo general:


Diseñar un clasificador de LTR retrotransposones en genomas de plantas utilizando métodos de *machine learning* y estrategias para incrementar su desempeño.

### OBJETIVOS ESPECÍFICOS:

- Seleccionar los algoritmos de *machine learning* para la clasificación de TEs.
- Establecer los hiperparámetros de los algoritmos de ML a evaluar.
- Evaluar el afinamiento de hiperparámetros con los rangos establecidos.
- Construir un algoritmo de ensamble con el fin de mejorar la precisión de los modelos experimentados.

## 7. METODOLOGÍA

**7.1. ORIGEN DE LOS DATOS:** Los datos usados en este proyecto fueron provistos por el tutor. Los elementos transponibles fueron obtenidos en primera instancia de 3 *sets* de datos ya conocidos: PGSB, Repbase and RepetDB. Adicionalmente, utilizando softwares de detección e identificación de TEs se obtuvieron otras muestras de elementos transponibles. Se aplicaron filtros a la base de datos

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

con el fin de conservar las secuencias intactas, sin deleciones o inserciones. La librería final no redundante contiene 67.305 elementos de 195 especies de plantas de 108 familias.

**7.2. CARACTERIZACIÓN DEL CONJUNTO DE DATOS:** debido a que el conjunto de datos son secuencias de ADN, estas son categóricas por lo que se debe realizar una transformación de los caracteres en representaciones numéricas. En este proyecto se utilizan frecuencia de k-mers como esquema de codificación, el cual consiste en registrar la cantidad de veces que se repite una combinación específica de nucleótidos. Las frecuencias de k-mers utilizadas para caracterizar el conjunto de datos fue usando valores de k desde 1 hasta 6, lo que genera 5.461 características.


**7.3. PRE-PROCESAMIENTO Y MÉTRICAS:** De acuerdo con Orozco-Arias (71), para este tipo de datos biológicos, los mejores desempeños se logran realizando escalamiento a los datos y aplicando el análisis de componentes principales conservando mínimo el 96% de varianza explicada, por tal motivo, al conjunto de datos en este proyecto se le realizó el pre-procesamiento sugerido. Adicionalmente, debido a que las clases se encuentran desbalanceadas, se utiliza como métrica del desempeño el f1-score (71) que es apropiada para la clasificación de elementos transponibles.

**7.5. DIVISION DE LOS DATOS:** EL conjunto de datos, se dividió en tres partes, 80% para entrenamiento, 10% para validación y 10% para prueba de los modelos, reduciendo así la posibilidad de sobre ajuste de los modelos.

**7.4. ENTRENAMIENTO DE LOS MODELOS:** Para el desarrollo del proyecto se utilizaron los modelos supervisados K-Neighbors Classifier (KNN), Logistic Regression (LR), Linear Support Vector Classifier (Linear SVC) y Linear Discriminant Analysis (LDA) los cuales fueron entrenados con sus valores por defecto y utilizando la estrategia de validación cruzada con el fin de obtener el desempeño base de los modelos con el f1-score.

**7.5. TUNEO DE HIPERPARÁMETROS:** Posterior al entrenamiento se indagaron los hiperparámetros de cada uno de los modelos que se muestran en la Tabla 1 y se establecieron los rangos de valores para los que iterar los algoritmos. Como estrategia para el tuneo de hiperparámetros se empleó el algoritmo GridSearchCV el cual se encarga de entrenar el clasificador con todas las combinaciones de valores de los hiperparámetros definidos y retorna los valores que alcanzan el mejor desempeño



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

para el modelo que se está tuneando. Adicionalmente, el algoritmo mide el desempeño usando la métrica ya establecida y el esquema de validación cruzada.


*Tabla 1. Modelos de ML, con los parámetros a tunear*

Classifier	Parameter	Range
<b>KNN</b>	neighbors	2,20,39,57,76,94,113,131,150
	weights	uniform, distance
	metric	euclidean, manhattan,chebyshev,minkowski,w minkowski,seuclidean,mahalanobis
	algorithm	auto, ball_tree,kd_tree,brute
<b>Linear SVC</b>	C	$1 \times 10^i$ con $i = -4$ , hasta 5
	penalty	l1, l2
	loss	hinge,squared_hinge
	tol	$10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}$
<b>LR</b>	C	$1 \times 10^i$ con $i = -4$ , hasta 5
	tol	$1 \times 10^i$ con $i = -4$ , hasta 5
	max_iter	$1 \times 10^i$ con $i = 0$ hasta 6
	penalty	l1, l2, elasticet,none
	solver	saga, liblinear,newton- cg,lbfgs,sag,saga
<b>LDA</b>	shrinkage	1,0.1,0.5,0.001,0.0001,0.00001
	solver	svd,lsqr,eigen
	tol	$10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}$

Este procedimiento se realizó con cada uno de los algoritmos establecidos en el inciso anterior teniendo en cuenta que algunos hiperparámetros no soportan valores de otros hiperparámetros. Para cada uno de los algoritmos se graficaron las curvas de aprendizaje del modelo con los mejores valores de hiperparámetros.

**7.6. MÉTODO DE ENSAMBLE:** Los métodos de ensamble de algoritmos son una agrupación de modelos de aprendizaje que clasifican nuevos datos tomando un ponderado de las predicciones de los modelos que lo conforman (72). Posterior al tuneo del conjunto de datos, se entrena el algoritmo



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

de ensamble con los métodos de mayor desempeño y se grafican las curvas de aprendizaje del modelo evaluándolo con el f1-score y utilizando la validación cruzada.

## 8. RESULTADOS

De acuerdo con la metodología planteada, para llevar a cabo el entrenamiento y tuneo de los modelos de *machine learning* se desarrollaron algoritmos en Python3 y fueron ejecutados en un servidor bajo el sistema operativo Linux con una capacidad de 52 procesadores y 256 Gb de RAM.

En el [Anexo I](#) se encuentra el algoritmo desarrollado para el procesamiento y ejecución de los análisis. Inicialmente se entrenaron los algoritmos establecidos en la metodología con sus parámetros por defecto para establecer un rendimiento base de los mismos con el conjunto de datos. Posteriormente, para cada uno de los modelos seleccionados se creó un diccionario que contenía el hiperparámetro y los valores a iterar. Luego de entrenar cada algoritmo de ML con el algoritmo *GridSearchCV* se obtuvieron los parámetros con los que se alcanzan los mejores desempeños y se relacionan en la Tabla 2. Después del tuneo de hiperparámetros se reentrenaron cada uno de los modelos para obtener su desempeño. Con el algoritmo LR se obtuvo un desempeño de 91% (Figura 1), Para KNN se obtuvo un desempeño del 97% (Figura 2), con LDA se obtuvo una precisión del 96% (Figura 3) y con Linear SVC se obtuvo un porcentaje de efectividad del 97% (Figura 4). En la Figura 5 se evidencian los desempeños de todos los algoritmos entrenados con los parámetros por defecto y los desempeños luego de afinarlos.

*Tabla 2. Valores de los parámetros tuneados.*

Classifier	Parameter	Value
KNN	neighbors	2
	weights	distance
	metric	euclidean
	algorithm	auto
Linear SVC	C	0.001
	penalty	l2
	loss	squared_hinge
	tol	0.1
LR	C	0.01
	tol	10

LDA	max_iter	1000
	penalty	l2
	solver	sag
	shrinkage	0.0001
	solver	lsqr
	tol	0.1

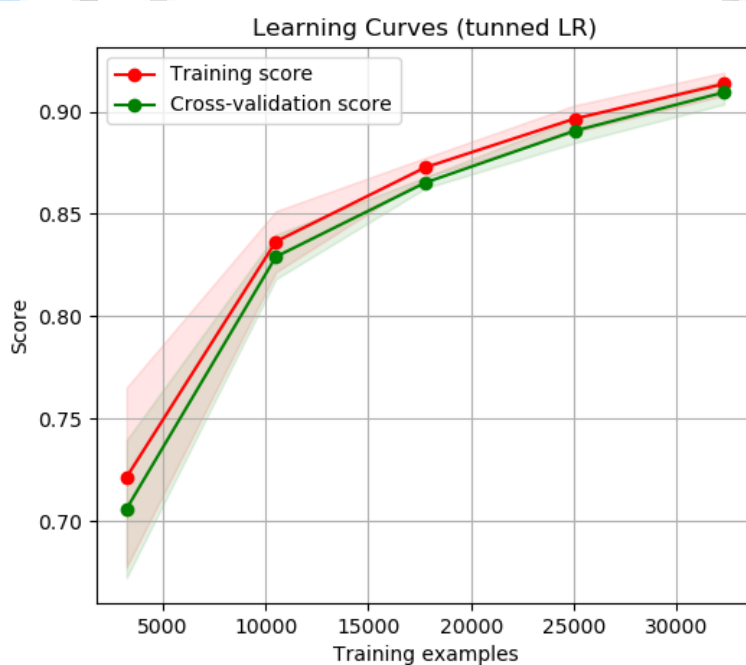


Figura 1. Curva de aprendizaje de modelo LR tuneado.

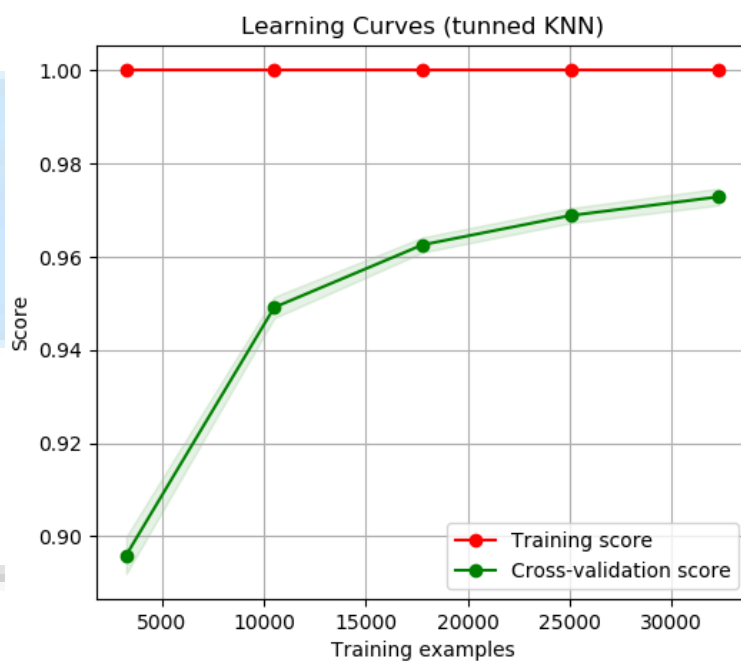


Figura 2. Curva de aprendizaje de modelo KNN tuneado.

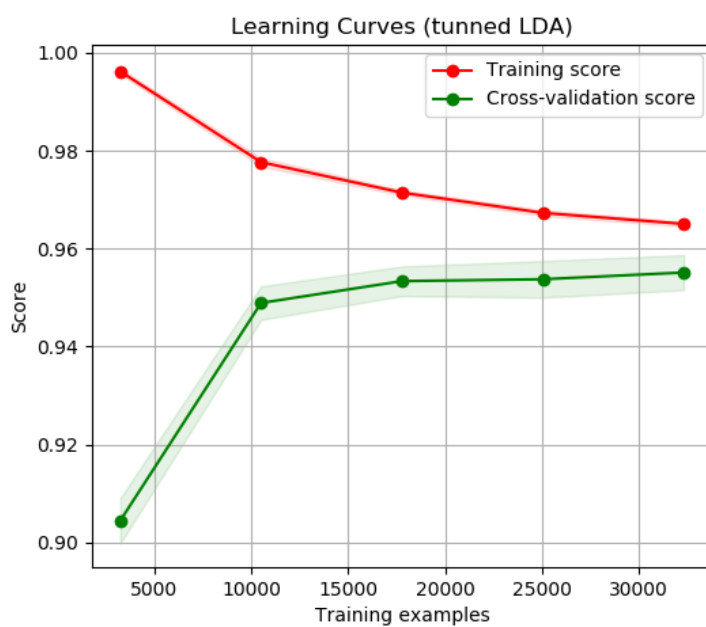


Figura 3. Curva de aprendizaje de modelo LDA tuneado.

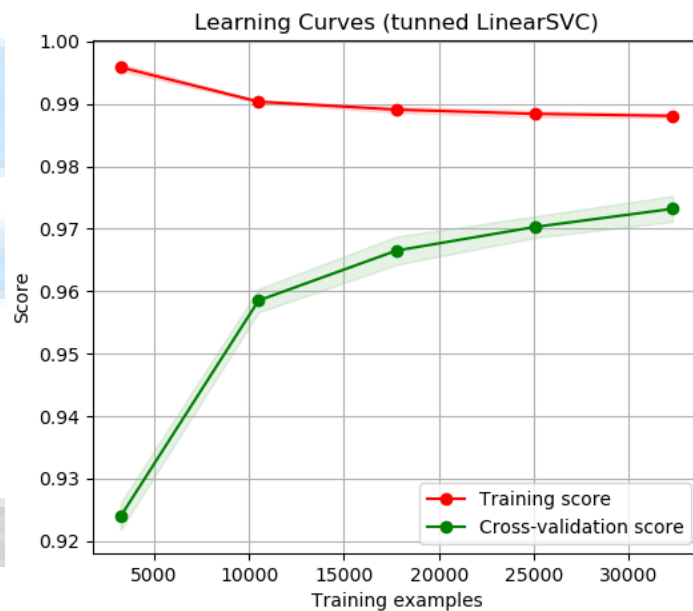


Figura 4. Curva de aprendizaje de modelo Linear SVC tuneado.

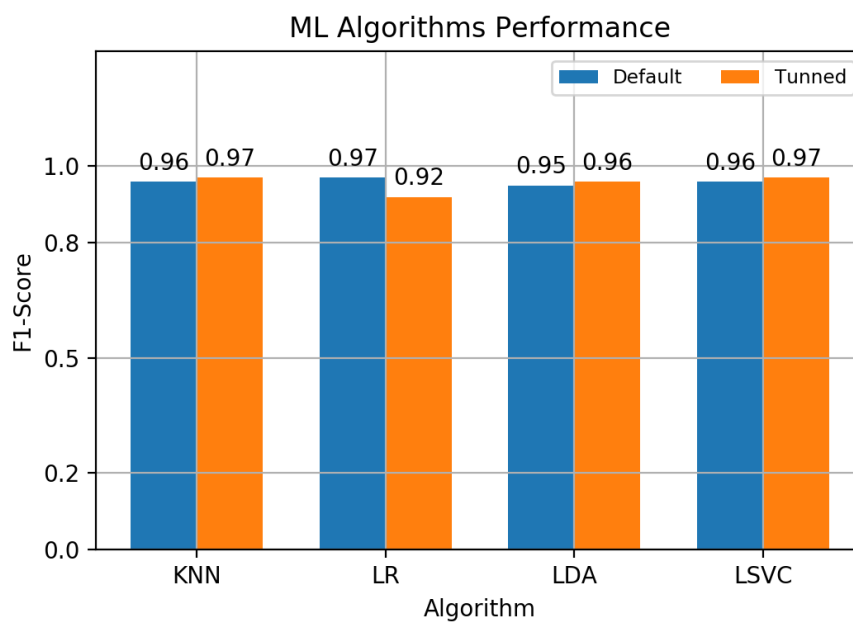


Figura 5. Desempeño de los algoritmos entrenados por defecto y los tuneados.

Para realizar el ensamble de algoritmos se excluyó el desempeño del clasificador LR ya que este presentaba el menor desempeño. Posteriormente se implementó como algoritmo de ensamble el *StackingClassifier* compuesto por los algoritmos LDA, Linear SVC y KNN tomando como meta clasificador el algoritmo *RandomForest*, el entrenamiento de este modelo de ensamble arrojó como resultado un desempeño del 97% (Figura 6) de efectividad para la clasificación de elementos transponibles. En la Figura 7 se puede observar el comportamiento de la clasificación del modelo de ensamble entrenado visualizando la matriz de confusión el cual entrega la información de la clasificación por cada una de las clases.

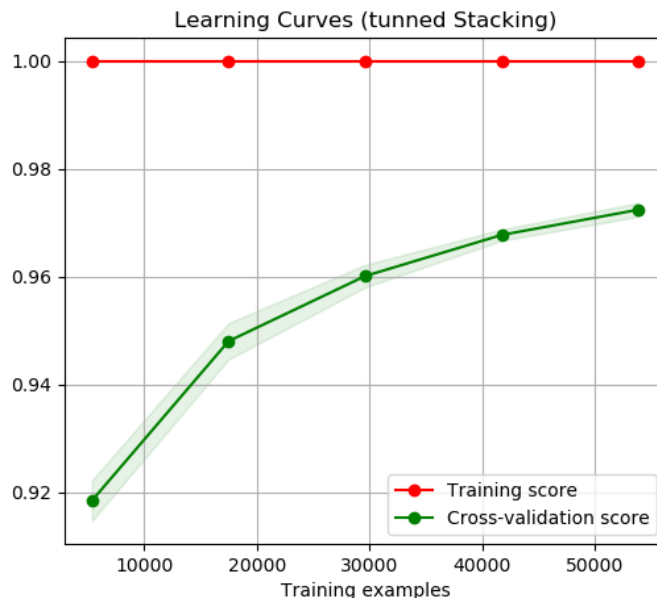


Figura 6. Curva de aprendizaje de modelo de ensamble.

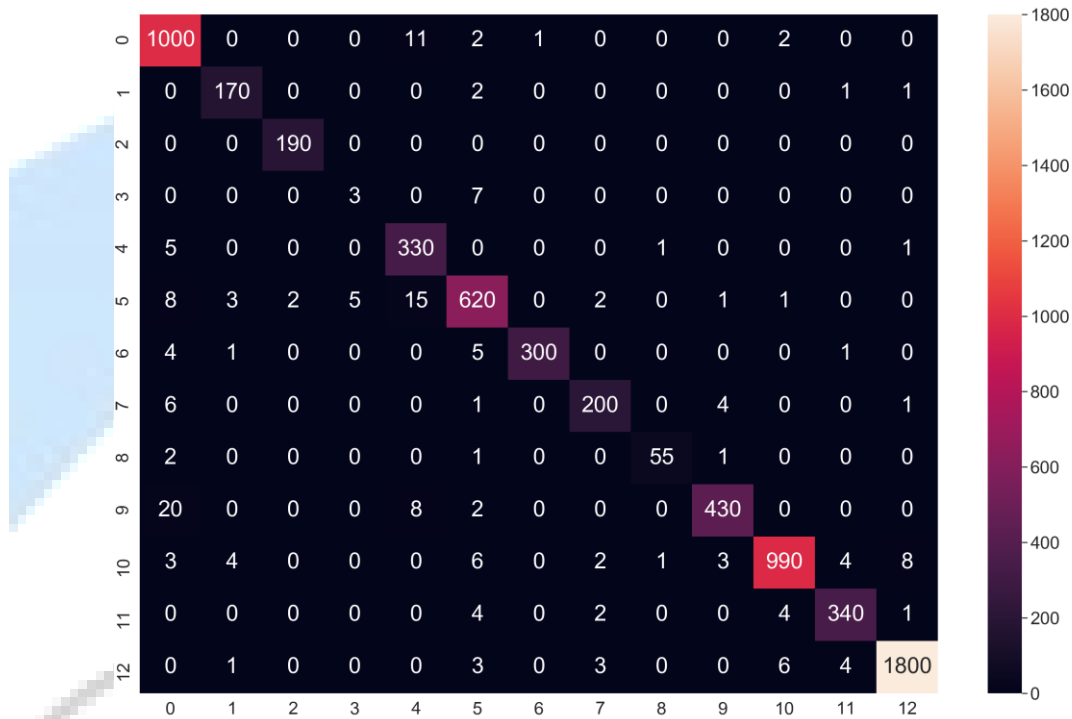



Figura 7. Matriz de confusión del algoritmo de ensamble entrenado

## 9. DISCUSIÓN DE RESULTADOS

De acuerdo a los resultados obtenidos en el proyecto se puede afirmar que los modelos de *machine learning* son efectivos para clasificar los elementos transponibles, especialmente los del orden LTR retrotransposones, logrando desempeños de hasta 97% de precisión. Con la ayuda de estrategias como el tuneo de hiperparámetros se puede alcanzar una mejora en desempeño de los algoritmos para clasificar los elementos transponibles. Aunque en términos de desempeño las mejoras no son significativas, ya que se logra mejorar hasta un 1% de precisión, las mejoras significativas se logran en cuanto a la generalización de los de los modelos ya que, al realizar la validación cruzada de los modelos, la desviación estándar que se logra está en el orden de  $10^{-3}$  permitiendo inferir que el modelo está realizando una correcta generalización y que las predicciones no están sesgadas por los datos de entrenamiento, o que los modelos propuestos tengan un buen balance entre flexibilidad y desempeño lo cual sugiere que emplear los modelos de *machine learning* descritos en este proyecto son una alternativa novedosa como técnica de clasificación de elementos transponibles

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

(específicamente LTR-Retrotransposones) a niveles muy altos de especificidad (linajes) disminuyendo así los falsos positivos y generando herramientas que permitan un entendimiento más profundo de estos elementos móviles y sus interacciones en los procesos genéticos que se llevan a cabo en las especies en las que están presentes, especialmente en las plantas de interés agrícola en Colombia y el mundo, haciéndolas resistentes a las condiciones climáticas agrestes ocasionadas por el cambio climático.

Desde el ámbito computacional, se pudo definir que los métodos de ensamble no mejoran significativamente el desempeño de los algoritmos por lo que se puede realizar un clasificador con los algoritmos tradicionales sin necesidad de recurrir a este tipo de técnicas que generan modelos con mayor complejidad. Aunque cabe destacar que las técnicas de ensamble ayudan a mejorar la generalización de los modelos de ML que lo componen, pero aumentan la complejidad.


## 10.CONCLUSIONES

Los resultados de la investigación permiten concluir que, de los modelos entrenados, los clasificadores con mejor desempeño fueron: linear SVC y KNN; Así mismo, el tuneo de hiperparámetros resulta una técnica muy útil para incrementar la efectividad de un modelo de *machine learning* con lo que la aplicación de estas estrategias para realizar la clasificación de los elementos transponibles resulta novedosa y presenta muy buenos resultados en cuanto a efectividad y velocidad. Los modelos de ensamble de algoritmo si bien son técnicas que pueden ayudar a mejorar el desempeño de los algoritmos de aprendizaje automático, en este proyecto se encontró que no aportan una mejora significativa al rendimiento en la clasificación de TEs.

## 11.RECOMENDACIONES

Como recomendaciones del proyecto se puede decir que para asegurar el correcto funcionamiento de los clasificadores se necesita aplicar pre-procesamiento a los datos con el fin de reducir su dimensionalidad y escalar los datos. Además, que cuando se realiza un tuneo de hiperparámetros se debe tener en cuenta que algunos hiperparámetros excluyen otras opciones de parámetros por lo que el tuneo deberá realizarse por etapas para evitar errores y fallos en los algoritmos.



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	CÓDIGO: GIN-GUI-001
		VERSIÓN: 01
		FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015

## 12.EVIDENCIA DE RESULTADOS EN GENERACIÓN DE CONOCIMIENTO, FORTALECIMIENTO DE LA CAPACIDAD CIENTÍFICA Y APROPIACIÓN SOCIAL DEL CONOCIMIENTO, FORMACIÓN

Relacionados con la generación de conocimiento y/o nuevos desarrollos tecnológicos:

Resultado	Indicador	Beneficiario
Artículo científico DOI: <a href="https://doi.org/10.3390/pr8060638">https://doi.org/10.3390/pr8060638</a>	Número de citas del artículo.	La Comunidad científica.

Conducentes al fortalecimiento de la capacidad científica nacional:

Resultado	Indicador	Beneficiario
Estudiante de pregrado formado a nivel profesional y con habilidades investigativas.	El presente informe de proyecto final.	Autor del proyecto, Universidad autónoma de Manizales.


Dirigidos a la apropiación social del conocimiento:

Resultado	Indicador	Beneficiario
Video de apropiación social del conocimiento.	Número de reproducciones en YouTube del video para medir el impacto de este.	Comunidad en general.

## 13. IMPACTOS LOGRADOS

Impacto esperado	Plazo (años) después de finalizado el proyecto: corto	Indicador verificable	Supuestos <sup>2</sup>


<sup>2</sup> Los supuestos indican los acontecimientos, las condiciones o las decisiones, necesarios para que se logre el impacto esperado.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

	(1-4), mediano (5-9), largo (10 o más)		
Identificación de los elementos transponibles de diversas especies de plantas de interés productivo en Colombia.	Corto plazo (1)	Porcentaje de efectividad obtenido en la clasificación de TEs con algoritmos tuneados de ML	Difusión de los resultados y continuidad de la línea de investigación.
Mover las barreras del conocimiento en el campo de los TEs y los procesos en los que están implicados	Corto Plazo (4)	Número de citas de los artículos publicados por el equipo de Bioinformática de la UAM relacionados.	Trabajo constante de los integrantes del equipo y asociación a los procesos en la UAM.


## 14. BIBLIOGRAFÍA

1. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes (Basel)* [Internet]. 2019 May 30 [cited 2020 Nov 27];10(6):419. Available from: <https://www.mdpi.com/2073-4425/10/6/419>
2. Sotero-Caio CG, Platt II RN, Suh A, Ray DA. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* [Internet]. 2017 Jan 1;9(1):161–77. Available from: <https://doi.org/10.1093/gbe/evw264>
3. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* [Internet]. 2017;18(2):71–86. Available from:

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


<https://doi.org/10.1038/nrg.2016.139>

4. Song MJ, Schaack S. Evolutionary conflict between mobile DNA and host genomes. *Am Nat.* 2018;192(2):263–73.
5. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. *Genet Res* [Internet]. 2009/04/14. 1983;42(1):1–27. Available from: <https://www.cambridge.org/core/article/population-dynamics-of-transposable-elements/B3815EB0AE7CDB71B27FDAE08A282B8A>
6. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* (80- ) [Internet]. 2014 Jul 18;345(6194):1249721. Available from: <http://science.sciencemag.org/content/345/6194/1249721.abstract>
7. Schietgat L, Vens C, Cerri R, Fischer CN, Costa E, Ramon J, et al. A machine learning based framework to identify and classify long terminal repeat retrotransposons. Bromberg Y, editor. *PLoS Comput Biol* [Internet]. 2018;14(4):e1006097. Available from: <https://dx.plos.org/10.1371/journal.pcbi.1006097>
8. Neumann P, Novák P, Hošťáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob DNA* [Internet]. 2019;10(1):1. Available from: <https://doi.org/10.1186/s13100-018-0144-1>
9. de Castro Nunes R, Orozco-Arias S, Crouzillat D, Mueller LA, Strickler SR, Descombes P, et al. Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions. *Front Plant Sci* [Internet]. 2018;9:175. Available from: <https://www.frontiersin.org/article/10.3389/fpls.2018.00175>
10. David T, Cruz S, Loureiro. Application of Machine Learning techniques on the Discovery and annotation of Transposons in genomes. In 2012.
11. Orozco Arias S, Isaza G, Guyot R. Retrotransposons in Plant Genomes: Structure,

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


Identification, and Classification through Bioinformatics and Machine Learning. *Int J Mol Sci.* 2019 Aug 6;20:1–31.

12. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics.* 2015 Jul;16.
13. Zytynicki M, Akhunov E, Quesneville H. Tedna: a transposable element de novo assembler. *Bioinformatics.* 2014 Sep;30(18):2656–8.
14. Staton SE, Burke JM. Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics [Internet].* 2015 Jun 1;31(11):1827–9. Available from: <https://doi.org/10.1093/bioinformatics/btv059>
15. Chu C, Nielsen R, Wu Y. REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. *PLoS One [Internet].* 2016 Mar 15;11(3):e0150719. Available from: <https://doi.org/10.1371/journal.pone.0150719>
16. Monat C, Tando N, Tranchant-Dubreuil C, Sabot F. LTRclassifier: A website for fast structural LTR retrotransposons classification in plants. *Mob Genet Elements [Internet].* 2016 Sep 26;6(6):e1241050–e1241050. Available from: <https://pubmed.ncbi.nlm.nih.gov/28090381>
17. Orozco-arias S, Liu J, Id RT, Ceballos D, Silva D, Id D, et al. Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology (Basel).* 2018;
18. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015 Jun;16(6):321–32.
19. Wassan J, Wang H, Zheng H. Machine Learning in Bioinformatics. In 2018.
20. Schietgat L, Vens C, Cerri R, Fischer CN, Costa E, Ramon J, et al. A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLoS Comput Biol [Internet].* 2018;14(4). Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046367727&doi=10.1371%2Fjournal.pcbi.1006097&partnerID=40&md5=7531a60dc9d1d9b0>


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

254ee4ab4fea87f2

21. Loureiro T, Camacho R, Vieira J, Fonseca NA. Boosting the Detection of Transposable Elements Using Machine Learning. Adv Intell Syst Comput [Internet]. 2013;222:85–91. Available from: [https://www.scopus.com/inward/record.uri?eid=2-s2.0-84880361824&doi=10.1007%2F978-3-319-00578-2\\_12&partnerID=40&md5=686e1fb2c8dacfac12b793834590d22d](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84880361824&doi=10.1007%2F978-3-319-00578-2_12&partnerID=40&md5=686e1fb2c8dacfac12b793834590d22d)
22. Alimentación | Naciones Unidas [Internet]. [cited 2020 Nov 27]. Available from: <https://www.un.org/es/sections/issues-depth/food/index.html>
23. El estado de la seguridad alimentaria y la nutrición en el mundo 2020. El estado de la seguridad alimentaria y la nutrición en el mundo 2020. FAO, IFAD, UNICEF, WFP and WHO; 2020.
24. Tito R, Vasconcelos HL, Feeley KJ. Global climate change increases risk of crop yield losses and food insecurity in the tropical Andes. Glob Chang Biol [Internet]. 2018;24(2):e592–602. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.13959>
25. Mustafin RN, Khusnutdinova EK. The Role of Transposons in Epigenetic Regulation of Ontogenesis. Russ J Dev Biol [Internet]. 2018;49(2):61–78. Available from: <https://doi.org/10.1134/S1062360418020066>
26. Bonchev G, Parisod C. Transposable elements and microevolutionary changes in natural populations. Mol Ecol Resour [Internet]. 2013 Sep 1;13(5):765–75. Available from: <https://doi.org/10.1111/1755-0998.12133>
27. Ata SK, Ou-Yang L, Fang Y, Kwoh C-K, Wu M, Li X-L. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. BMC Syst Biol [Internet]. 2018;12(S9):138. Available from: <https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0662-y>
28. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: Where genetics meets genomics. Nat Rev Genet. 2002;3(5):329–41.


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

29. Li S-F, Su T, Cheng G-Q, Wang B-X, Li X, Deng C-L, et al. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes (Basel)* [Internet]. 2017 Oct 24 [cited 2020 Nov 28];8(10):290. Available from: <http://www.mdpi.com/2073-4425/8/10/290>
30. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. *Mol Ecol*. 2013;22(6):1503–17.
31. Heo GE, Kang KY, Song M, Lee J-H. Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. *BMC Bioinformatics* [Internet]. 2017 May;18(S7):251. Available from: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1640-x>
32. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An Automatic Transposable Element Classification Tool. Cordaux R, editor. *PLoS One* [Internet]. 2014 May 2 [cited 2020 Nov 28];9(5):e91929. Available from: <https://dx.plos.org/10.1371/journal.pone.0091929>
33. Bousios A, Minga E, Kalitsou N, Pantermali M, Tsaballa A, Darzentas N. MASiVEDb: the Sirevirus Plant Retrotransposon Database. *BMC Genomics* [Internet]. 2012;13(1):158. Available from: <https://doi.org/10.1186/1471-2164-13-158>
34. Abrusan G, Grundmann N, DeMester L, Makalowski W. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *BIOINFORMATICS*. 2009 May;25(10):1329–30.
35. Abrusán G, Grundmann N, Demester L, Makalowski W. TEclass - A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* [Internet]. 2009;25(10):1329–30. Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-65549101705&doi=10.1093%2Fbioinformatics%2Fbtp084&partnerID=40&md5=540916e071ea e729d30cc45600aa708f>
36. Dashti TH, Masoudi-Nejad A. Mining Biological Repetitive Sequences Using Support Vector Machines and Fuzzy SVM. *Iran J Chem Chem Eng ENGLISH Ed*. 2010;29(4):1–17.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


37. Li Q, Zhang Y, Zhang Z, Li X, Yao D, Wang Y, et al. A D-genome-originated Ty1/Copia-type retrotransposon family expanded significantly in tetraploid cottons. *Mol Genet Genomics* [Internet]. 2018;293(1):33–43. Available from: <https://doi.org/10.1007/s00438-017-1359-4>
38. Schulman AH. Retrotransposon replication in plants. *Curr Opin Virol* [Internet]. 2013;3(6):604–14. Available from: <http://www.sciencedirect.com/science/article/pii/S1879625713001454>
39. Negi P, Rai AN, Suprasanna P. Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. *Front Plant Sci* [Internet]. 2016;7:1448. Available from: <https://www.frontiersin.org/article/10.3389/fpls.2016.01448>
40. Kejnovsky E, Tokan V, Lexa M. Transposable elements and G-quadruplexes. *Chromosom Res* [Internet]. 2015;23(3):615–23. Available from: <https://doi.org/10.1007/s10577-015-9491-7>
41. Hermann D, Egue F, Tastard E, Nguyen DH, Casse N, Caruso A, et al. An introduction to the vast world of transposable elements - What about the diatoms? *Diatom Res* [Internet]. 2014;29(1):91–104. Available from: <https://doi.org/10.1080/0269249X.2013.877083>
42. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, et al. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol*. 2010 Apr;186(1):37–45.
43. Lyon MF. LINE-1 elements and X chromosome inactivation: a function for “junk” DNA? *Proc Natl Acad Sci U S A* [Internet]. 2000 Jun 6;97(12):6248–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/10841528>
44. Kim N-S. The genomes and transposable elements in plants: are they friends or foes? *Genes Genomics* [Internet]. 2017;39(4):359–70. Available from: <https://doi.org/10.1007/s13258-017-0522-y>
45. Vicient CM, Casacuberta JM. Impact of transposable elements on polyploid plant genomes. *Ann Bot* [Internet]. 2017 Aug 1;120(2):195–207. Available from: <https://doi.org/10.1093/aob/mcx078>
46. Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. A computational study of the dynamics of



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


LTR retrotransposons in the *Populus trichocarpa* genome. *Tree Genet Genomes* [Internet]. 2012;8(1):61–75. Available from: <https://doi.org/10.1007/s11295-011-0421-3>

47. Ferreira de Carvalho J, Chelaifa H, Boutte J, Poulain J, Couloux A, Wincker P, et al. Exploring the genome of the salt-marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC end sequence analysis. *Plant Mol Biol* [Internet]. 2013;83(6):591–606. Available from: <https://doi.org/10.1007/s11103-013-0111-7>
48. Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. *Tree Genet Genomes* [Internet]. 2017;13(5):96. Available from: <https://doi.org/10.1007/s11295-017-1181-5>
49. Paz RC, Kozaczek ME, Rosli HG, Andino NP, Sanchez-Puerta MV. Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in *Solanum lycopersicum*. *Genetica* [Internet]. 2017;145(4):417–30. Available from: <https://doi.org/10.1007/s10709-017-9977-7>
50. Griffiths J, Catoni M, Iwasaki M, Paszkowski J. Sequence-Independent Identification of Active LTR Retrotransposons in *Arabidopsis*. *Mol Plant*. 2018;11(3):508–11.
51. Wang H, Huang H, Ding C. Function-Function Correlated Multi-Label Protein Function Prediction over Interaction Networks. In 2012. p. 302–13. Available from: [http://link.springer.com/10.1007/978-3-642-29627-7\\_32](http://link.springer.com/10.1007/978-3-642-29627-7_32)
52. Fan F, Wen X, Ding G, Cui B. Isolation, identification, and characterization of genomic LTR retrotransposon sequences from masson pine (*Pinus massoniana*). *Tree Genet Genomes* [Internet]. 2013;9(5):1237–46. Available from: <https://doi.org/10.1007/s11295-013-0631-y>
53. Rawal K, Ramaswamy R. Genome-wide analysis of mobile genetic element insertion sites. *Nucleic Acids Res* [Internet]. 2011 Sep;39(16):6864–78. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr337>
54. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* [Internet]. 2017;8(1):19. Available from:


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

<https://mobile.dnajournal.biomedcentral.com/articles/10.1186/s13100-017-0103-2>

55. Jiang S-Y, Ramachandran S. Genome-Wide Survey and Comparative Analysis of LTR Retrotransposons and Their Captured Genes in Rice and Sorghum. PLoS One [Internet]. 2013 Jul 29;8(7):e71118. Available from: <https://doi.org/10.1371/journal.pone.0071118>
56. Loureiro T, Camacho R, Vieira J, Fonseca NA. Improving the performance of Transposable Elements detection tools. J Integr Bioinform [Internet]. 2013 Nov;10(3):231. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24231145>
57. Jiang N. Overview of Repeat Annotation and De Novo Repeat Identification BT - Plant Transposable Elements: Methods and Protocols. In: Peterson T, editor. Totowa, NJ: Humana Press; 2013. p. 275–87. Available from: [https://doi.org/10.1007/978-1-62703-568-2\\_20](https://doi.org/10.1007/978-1-62703-568-2_20)
58. Nicolas J, Peterlongo P, Tempel S. Finding and Characterizing Repeats in Plant Genomes BT - Plant Bioinformatics: Methods and Protocols. In: Edwards D, editor. New York, NY: Springer New York; 2016. p. 293–337. Available from: [https://doi.org/10.1007/978-1-4939-3167-5\\_17](https://doi.org/10.1007/978-1-4939-3167-5_17)
59. Chiusano ML, Colantuono C. Repeat Sequences in the Tomato Genome BT - The Tomato Genome. In: Causse M, Giovannoni J, Bouzayen M, Zouine M, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 173–99. Available from: [https://doi.org/10.1007/978-3-662-53389-5\\_10](https://doi.org/10.1007/978-3-662-53389-5_10)
60. Pang E, Cao H, Zhang B, Lin K. Crop Genome Annotation: A Case Study for the Brassica rapa Genome BT - The Brassica rapa Genome. In: Wang X, Kole C, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015. p. 53–64. Available from: [https://doi.org/10.1007/978-3-662-47901-8\\_5](https://doi.org/10.1007/978-3-662-47901-8_5)
61. Smit AF. Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res [Internet]. 1993 Apr 25;21(8):1863–72. Available from: <https://pubmed.ncbi.nlm.nih.gov/8388099>
62. Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res. 2008;134(1–2):221–34.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

63. Arango-López J, Orozco-Arias S, Salazar JA, Guyot R. Application of data mining algorithms to classify biological data: The coffea canephora genome case. Commun Comput Inf Sci [Internet]. 2017;735:156–70. Available from: [https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028831337&doi=10.1007%2F978-3-319-66562-7\\_12&partnerID=40&md5=8b54eab4c9753fe6c2fb5037a930c1bc](https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028831337&doi=10.1007%2F978-3-319-66562-7_12&partnerID=40&md5=8b54eab4c9753fe6c2fb5037a930c1bc)
64. Mjolsness E, Decoste D. Machine Learning for Science: State of the Art and Future Prospects. Science. 2001;293:2051–5.
65. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet [Internet]. 2019;51(1):12–8. Available from: <https://doi.org/10.1038/s41588-018-0295-5>
66. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci [Internet]. 2014;19(12):798–808. Available from: <http://www.sciencedirect.com/science/article/pii/S1360138514002192>
67. Arango-López J, Orozco-Arias S, Salazar JA, Guyot R. Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case BT - Advances in Computing. In: Solano A, Ordoñez H, editors. Cham: Springer International Publishing; 2017. p. 156–70.
68. Ashlock W, Datta S. Distinguishing endogenous retroviral LTRs from SINE elements using features extracted from evolved side effect machines. IEEE/ACM Trans Comput Biol Bioinforma [Internet]. 2012;9(6):1676–89. Available from: <http://dx.doi.org/10.1109/TCBB.2012.116>
69. Nakano FK, Pinto WJ, Pappa GL, Cerri R. Top-down strategies for hierarchical classification of transposable elements with neural networks. In: Proceedings of the International Joint Conference on Neural Networks [Internet]. 2017. p. 2539–46. Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030973170&doi=10.1109%2FIJCNN.2017.7966165&partnerID=40&md5=be2dead241b3e3326d7f53fab0fe99f>
70. Garbus I, Romero JR, Valarik M, Vanžurová H, Karafiátová M, Cáccamo M, et al.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. BMC Genomics [Internet]. 2015;16(1):375. Available from: <https://doi.org/10.1186/s12864-015-1579-0>

71. Orozco-Arias S, Piña JS, Tabares-Soto R, Castillo-Ossa LF, Guyot R, Isaza G. Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. Processes. 2020;8(6):1–20.
72. Dietterich TG. Ensemble methods in machine learning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2000;1857 LNCS:1–15.

## 15.ANEXOS

### ANEXO 1. EVIDENCIA DE LOS CODIGOS FUENTES QUE SE DESARROLLARON DURANTE EL PROYECTO.