**PROYECTO FORMATIVO PARA PROGRAMACIÓN EN ANALITICA DE DATOS – TPAD**

**FICHA:** 2067782

**INTEGRANTES:**

Alejandra Rentería Tenorio

Chadia Vanessa Jiménez

Hugo Santacruz

Jean Paul Benítez

Juan Pablo Serrano Rodríguez

Leidy Katerine Benavidez

Leidy Paola Rodallega

Paula Valentina Vargas

Santiago Giraldo

Tatiana Nieto



Servicio Nacional de Aprendizaje - SENA

Centro de Electricidad y Automatización Industrial- CEAI

Santiago de Cali, Febrero xx 2021

REPÚBLICA DE COLOMBIA

**PROYECTO FORMATIVO PARA PROGRAMACIÓN EN ANALITICA DE DATOS – TPAD**

**FICHA:** 2067782

Análisis BIG DATA de la producción del café en Colombia utilizando técnicas y procedimientos de los lenguajes Python y Rstudio para la construcción de un modelo de regresión lineal (**CAFELÍTICA**)

**INTEGRANTES:**

Alejandra Rentería Tenorio

Chadia Vanessa Jiménez

Hugo Santacruz

Jean Paul Benítez

Juan Pablo Serrano Rodríguez

Leidy Katerine Benavidez

Leidy Paola Rodallega

Paula Valentina Vargas

Santiago Giraldo

Tatiana Nieto

Servicio Nacional de Aprendizaje - SENA

Centro de Electricidad y Automatización Industrial – CEAI

Santiago de Cali, Febrero xx 2021

REPÚBLICA DE COLOMBIA

**CONTENIDO**

[1. Introducción y Resumen 7](#_Toc63793258)

[2. Título del proyecto 10](#_Toc63793259)

[3. Planteamiento del problema o necesidad que se pretende solucionar 10](#_Toc63793260)

[4. Justificación del proyecto 13](#_Toc63793261)

[5. Objetivo General 15](#_Toc63793262)

[5.1. Objetivos Específicos 16](#_Toc63793263)

[6. Alcance del proyecto 16](#_Toc63793264)

[7. Beneficiarios del Proyecto. 17](#_Toc63793265)

[8. Impactos del Proyecto 18](#_Toc63793266)

[8.1. Económico 18](#_Toc63793267)

[8.2. Tecnológico 18](#_Toc63793268)

[9. Restricciones o riesgos asociados y alternativas de solución 20](#_Toc63793269)

[9.1 Restricción de Costo 20](#_Toc63793270)

[9.2 Restricción de Tiempo 21](#_Toc63793271)

[9.3 Restricción de Alcance 23](#_Toc63793272)

[10. Productos o resultados del desarrollo del proyecto 24](#_Toc63793273)

[10.1 Metodología Empleada 24](#_Toc63793274)

[10.2 Requerimientos de los datos 27](#_Toc63793276)

[10.3 Entendimiento de los datos 28](#_Toc63793277)

[10.4 Preparación de los datos 29](#_Toc63793278)

[10.5 Preparación del modelo 31](#_Toc63793279)

[10.6 Evaluación del Modelo 33](#_Toc63793280)

[11. Entregables y su descripción 33](#_Toc63793281)

[12. Conclusiones 34](#_Toc63793282)

[12.1 Bibliografía 34](#_Toc63793283)

# Tabla De Gráficas

[Tabla 1, Cotización de la materia prima 20](#_Toc64052269)

[Tabla 2, Notas del mes Septiembre 22](#_Toc64052270)

[Tabla 3, Notas del mes Octubre 22](#_Toc64052271)

[Tabla 4, Notas del mes Noviembre 23](#_Toc64052272)

[Tabla 5, Diagnostico estadístico de la mensual del café Colombiano 31](#_Toc64052273)

# Tabla De Ilustraciones

[Ilustración 1, Factor de rendimiento en Colombia, Fuente FNC 12](#_Toc64052190)

[Ilustración 2, Defectos del café y sus orígenes. Fuente: FNC 13](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052191)

[Ilustración 3, Metodología Crips-Dm. 26](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052192)

[Ilustración 4, Mapa de calor que relaciona las variables utilizando la librería Seaborn 27](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052193)

[Ilustración 5, Histogramas y polígonos entre el rendimiento, área y producción nacional 28](#_Toc64052194)

[Ilustración 6, Dispersión del Rendimiento del café acumulado por año 29](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052195)

[Ilustración 7, Rendimiento por Area nacional en secuencia de los años 29](#_Toc64052196)

[Ilustración 13, Modelo de Pearson 33](#_Toc64052197)

[Ilustración 14, Modelo Spearman 33](#_Toc64052198)

[Ilustración 15, Modelo Kendall 34](#_Toc64052199)

[Ilustración 16, Eliminación de valores faltantes en Python 35](#_Toc64052200)

[Ilustración 17, Conteo de valores duplicados en Python 35](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052201)

[Ilustración 18, Inspección de valores nulos en R 36](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052202)

[Ilustración 19, Columna de valores perdidos 36](#_Toc64052203)

[Ilustración 20, Visualización del modelo lineal en Python 37](file:///D:\Familia\Downloads\Informe%20proyecto%20formativo%20mezcla%20TPAD1-V10.docx#_Toc64052204)

[Ilustración 21, Grafico de dispersión Área N vs Rendimiento en RStudi 38](#_Toc64052205)

[Ilustración 22, Modelo de regresión lineal Área n. vs Rendimiento en RStudio 38](#_Toc64052206)

[Ilustración 23, Grafica estimación del Área Nacional cosechada por el Rendimiento del café 39](#_Toc64052207)

# Introducción y Resumen

El café no es solo un producto agrícola de exportación, sigue siendo una de las bebidas más consumidas en el mundo, contribuyendo a la estabilidad e integración económica de Colombia, permitiéndole existir en el mercado internacional por mucho tiempo. Lo que hace famoso al café no son solo sus características emblemáticas sino su producción, exportación, calidad y variedad que lo han posicionado de manera favorable en los últimos tiempos.

La actividad cafetera significa para el país: desarrollo rural, empleo, ingresos y durabilidad. Su fuerte impacto e influencia en la economía de Colombia lo ha posicionado en lo alto por demostrar su eficiencia en la producción del café manteniéndolo al margen a pesar de las diferentes crisis. Actualmente, Colombia ocupa el tercer lugar en las exportaciones de café del mundo, después de Brasil y Vietnam. Cabe resaltar, que hoy el café sigue siendo la gran entrada de divisas del país y alimenta la economía de diferentes sectores de la población Colombiana, desde los que cultivan granos hasta los responsables de las exportaciones. Demostrando que el café es cultura propia de los colombianos donde se entrelaza con la economía, la política y demás. Este producto es considerado denominación de origen.

El presente proyecto tiene como objetivo observar el comportamiento, así mismo el rendimiento, la producción y el área de cultivo a nivel nacional que ha presentado el café colombiano en sus diferentes departamentos a través de los años. Con el apoyo de diversas fuentes se logra crear un proceso de inspección de atributos que posee el fruto.

Sabiendo que al ser tan importante para los consumidores debe pasar a ser investigado por científicos capaces de obtener mejoras y sugerir cambios decisivos en la calidad del café.

A lo largo del informe encontraremos temas claves para el desarrollo del proyecto como las problemáticas que presenta la producción del café, las técnicas y herramientas utilizadas para la ejecución del BIG DATA que permitan solucionar los contratiempos que se puedan manifestar y favorezcan a los productores del café, el impacto y las restricciones que causa el proceso de la preparación de los datos.

Se evaluará la producción del café en Colombia donde se hace una comparación frente a los últimos años de producción, buscando así mejorar su rendimiento. Cabe señalar, que a lo largo de los años, muchas crisis fuertes han traído complejidad y desventajas a la producción de café, no solo estamos hablando de la crisis económica, sino también ambiental, climática, la falta de inversión, las plagas, la competencia internacional y la falta de rendimiento, etc. Para esto se ha optado por realizar un estudio de análisis de datos estadísticos bajo el dominio de herramientas computacionales de análisis que faciliten una rápida absorción de la información, para evaluar una muestra específica, permitiendo que nos revele relaciones o dependencias entre variables indicando los puntos clave a tratar para ser implementados a la hora de pensar en el café como un producto esencial y de gran importancia para la mejora de la economía del país , pues no hay duda, de que si se logra un excelente desempeño en producción y rendimiento, se lograría volver a ocupar hasta el primer lugar de países exportadores de café, dirigiendo al país a un fuerte progreso, para su ejecución se emplearán técnicas de Machine Learning para la predicción del rendimiento del café, contribuyendo al avance del sector, generando nuevo conocimiento, brindando herramientas para futuras investigaciones y aportando conclusiones que sirvan de base para proponer estrategias que permitan mejorar la competitividad del sector.

# **Título del proyecto**

ANÁLISIS BIG DATA DE LA PRODUCCIÓN DEL CAFÉ EN COLOMBIA UTILIZANDO TÉCNICAS Y PROCEDIMIENTOS DE LOS LENGUAJES PYTHON Y R PARA LA CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LINEAL (**CAFELÍTICA**)

# Planteamiento del problema o necesidad que se pretende solucionar

La problemática que se aborda en el presente proyecto, es la producción, el rendimiento, el área sembrada del café y la manera que se comercializa reflejando sus costos de producción altos y el ingreso que reciben por las ventas demasiado bajo. En ocasiones, este ingreso no es suficiente ni siquiera para cubrir los costos de producción y mucho menos para dejarle una ganancia al pequeño productor de café, debido a las distorsiones del comportamiento del mercado como lo son las cotizaciones en la Bolsa de **Nueva York Commodity Exchange**. A pesar de ello este precio vacila regularmente, ocasionando que se gaste más de lo que se gana. Y esta termina siendo la causante de que los bajos precios suelan inducir a los productores, en particular a los menos competitivos, a reducir la producción de sus cultivos.

La bolsa de cotizacionesprima por la calidad que se le reconoce al café colombiano y la tasa de cambio del peso colombiano frente al dólar.

Según la Federación Nacional de Caficultores (FNC), el factor de rendimiento es la cantidad de café pergamino seco (CPS) que se necesita para obtener un saco de 70 kilos de café excelso y se determina al momento de realizar el análisis físico del café que se va a vender.

Dadas las condiciones de calidad del café en nuestro país, el promedio del factor de rendimiento es de 92.8 es decir, que se necesitan 92,8 kilos de café pergaminos seco para obtener 70 kilos de excelso (grano de café bueno).

Al momento de vender el café en el punto de compra se tiene en cuenta, **entre menor factor de rendimiento tenga el café, mayor es el precio que obtendrá el productor.**

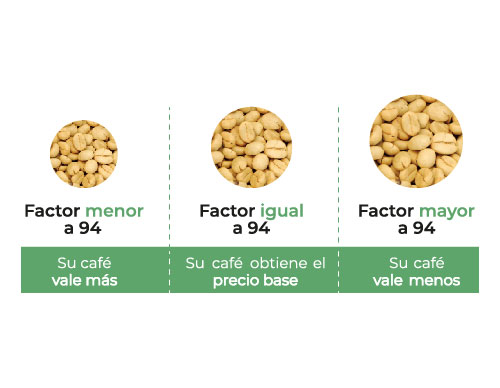


Ilustración 1, Factor de rendimiento en Colombia, Fuente FNC https://federaciondecafeteros.org/wp/servicios-al-caficultor/aprenda-a-vender-su-cafe

El mercado del café es un ejemplo de crisis estructural que enfrenta un mercado globalizado, que coloca a los productores de café en una situación de inseguridad en el futuro, mientras que las áreas de producción enfrentan serias dificultades de desarrollo.

En el campo de la agricultura se han elaborados estudios por medio del uso de técnicas de aprendizaje automatizado, con el fin de predecir las mejores condiciones para los cultivos analizados. Algunos se centran en factores climáticos, composición del suelo o incluso el uso de agentes químicos que contrarrestan algún patógeno del cultivo o potencian su crecimiento, así pueden contribuir con programas e investigaciones que permitan a los caficultores adquirir conocimientos y aplicar nuevas tecnologías para incrementar el rendimiento de los cultivos y productividad del grano.

Para complementar las problemáticas, se encuentra que al manejar una gran área de cosecha, al momento de cultivar el café no se tienen en cuenta el porcentaje que es recogido. Por lo que desconocerán cuales son las ganancias, y encima, se encuentra imposible que el personal logre observar y proteger a las plantaciones de los diferentes fenómenos naturales que suceden en las áreas donde se cultiva.

El siguiente gráfico profundiza los efectos que tiene ciertos fenómenos en el proceso del cultivo que alteran su apariencia física, afectan el sabor y el aroma del grano.



Ilustración 2, Defectos del café y sus orígenes. Fuente: FNC, https://federaciondecafeteros.org/wp/servicios-al-caficultor/aprenda-a-vender-su-cafe/ (último acceso: 28 de 01 de 2021).

Por lo anterior se tiene la pregunta:

**¿Cómo predecir el rendimiento de la producción del café en Colombia, mediante un modelo analítico en Python, RStudio que utilice las técnicas y herramientas proporcionadas por BIG DATA?**

# Justificación del proyecto

Ante los enormes retos de estrategia y gestión de recursos que componen el proceso de toma de decisiones de las empresas agrícolas, estas hacen uso de las TIC para apoyar a los problemas que enfrentan a diario. Un ejemplo, son las actividades cafetaleras que procesan una gran cantidad de datos, pero que no saben cómo gestionarlos de manera eficaz, útil y rápida para comprender mejor los factores de producción y rendimiento en el mercado, así como realizar planes y programas estratégicos. Para identificar estos patrones, fue importante la realización del análisis de los datos recolectados durante los años 2007-2018 sobre el cultivo y la producción del café en Colombia, para poder entender y comprender los factores que impactan el comportamiento de una variable, como lo es el rendimiento del café, porque según la actividad agrícola tradicional esta se basa en la intuición y la experiencia en tareas y calendarios específicos de siembra.

Por eso es esencial que los caficultores no sigan tomando decisiones según su intuición y permitan implementar las tecnologías de Big data para que faciliten la toma de decisiones y se basen en las cifras en cuanto a producción y rendimiento de los cultivos y demás. De acuerdo a lo anterior, el análisis de datos juega un papel importante que les permitirá aumentar la fiabilidad de sus operaciones y reducir el riesgo de pérdida en sus inversiones teniendo mejores rendimientos en sus cultivos, haciéndolos rentables porque adoptarían las mejores prácticas existentes, teniendo como base el conocimiento de las tendencias y/o patrones identificados en los datos.

La tarea de monitorear macro datos (Big Data), acerca del café es importante, ya que permitirá visualizar y manipular la información para crear tablas, gráficas y modelos predictivos, basados en métodos, formularios y análisis estadísticos estandarizados, dándole una mejor supervisión de lo que sucede en la superficie de la tierra en cada momento, optimizando en términos de costos y producción las diferentes zonas agroclimáticas a partir del uso de tecnologías digitales, incluso dar a conocer los efectos adversos que puede tener una variable impredecible sobre la cosecha.

Este proyecto es fundamental porque a través de su formulación y desarrollo permitirá predecir comportamientos en la producción y el rendimiento del café en Colombia utilizando técnicas y herramientas propias de la ciencia de datos, Big Data, aprendizaje automático, etc., desarrollando un sistema de soporte que mejore la precisión de la gestión de cultivos en entornos de incertidumbre y fortalezca el proceso eficaz y eficiente de la toma de decisiones.

Dentro de algunos de los aportes relacionados con las técnicas de aprendizaje automático podemos mencionar la siguiente:

* Las técnicas de aprendizaje automático permiten que una máquina sea capaz de aprender, reconocer patrones, características y vínculos entre un conjunto de datos, que muchas veces están dentro del dominio del conocimiento tácito de expertos o artesanos dedicados a una labor específica. Este aprendizaje se logra mediante el uso de datos de entrenamiento, que relacionan unos datos de entrada con unos resultados de salida. En el mundo se vienen desarrollando diferentes investigaciones…, en donde la complejidad o la especialidad de una tarea, representan un desafío muy grande de programación con las técnicas tradicionales, convirtiendo al MACHINE LEARNING en una herramienta que permite emular dichos conocimientos o técnicas y facilitar el uso de máquinas para la realización de tareas complejas. Descripción: Comillas, Símbolo, Diseño De Iconos imagen png - imagen transparente  descarga gratuita (Peña, 2019)
* Por otro lado, gracias a los aportes de la inteligencia artificial y al machine learning en los diferentes campos mencionados se ha hecho posible hacer la vida del ser humano más fácil, permitiendo el progreso económico, tecnológico y social en todos los ámbitos de nuestra existencia. Descripción: Comillas, Símbolo, Diseño De Iconos imagen png - imagen transparente  descarga gratuita (Ramírez, 2018).

# Objetivo General

El objetivo principal es contribuir al desarrollo de un modelo predictivo mediante herramientas computacionales de análisis (Python o Rstudio) que nos permita mejorar el factor de rendimiento en la producción del café en Colombia de acuerdo a las tendencias y/o patrones identificados, para así incorporar técnicas de Big Data que ayude a los caficultores a supervisar en tiempo real los datos recopilados a la hora de tomar decisiones según los niveles de productividad y rendimiento que se obtengan.

La finalidad del proyecto es que al usar las tecnologías y las herramientas provistas por la ciencia de datos para el análisis del Big Data en la agricultura, se logré construir un modelo predictivo automático utilizando los lenguajes tales como Python y R que permita determinar la producción o el rendimiento del café en Colombia.

## Objetivos Específicos

* Realizar la recolección y búsqueda de los datos sobre el proceso del cultivo del café en Colombia
* Utilizar un software como RStudio y Python para realizar el proceso de ETL sobre los datos
* Crear un modelo predictivo basado en aprendizaje automático utilizando técnicas y herramientas de Python y R que permita determinar el rendimiento del café en Colombia
* Evaluar la efectividad del modelo de aprendizaje automático alimentándolo con variables y muestras de la producción y rendimiento del café en Colombia.

# Alcance del proyecto

Para alcanzar la meta se debe encontrar la forma de aumentar el índice de rendimiento en la producción del café en Colombia, identificando los factores que determinarán las acciones que se deben implementar con decisiones apoyadas en la información y en herramientas de análisis de datos. En este orden de ideas, es necesario profundizar en los factores que influyen a este indicador productivo mediante la recolección de información de los cultivos de café en Colombia donde el resultado esperado consiste en un modelo predictivo, que represente el comportamiento correlacionado de las variables de rendimiento con las variables estadísticas de producción. Llevándose a cabo la recolección de información a nivel nacional, en donde se realizarán los análisis previos (Análisis de Diagnostico, Análisis Descriptivo y Análisis Predictivo). En estos datos analizados se pretende eliminar los errores en la información, mediante procesos de unificación de los datos para el caso de duplicidad e imputación de datos con el fin de mejorar su calidad y estructura, permitiendo mitigar así los riesgos de error en la información y finalmente crear un informe que sea entendible a simple vista y a su vez permita generar cambios en sus decisiones.

# Beneficiarios del Proyecto.

Los principales beneficiarios serán los caficultores que obtendrán el acceso a nuevas herramientas para la toma de decisiones y así puedan tener un crecimiento organizacional, económico y tecnológico, enseñándoles a los sectores agrícolas a extraer el máximo beneficio económico de los datos. Y lo hace de un modo muy simple: acercando a la gente de los datos a la cultura cafetera y viceversa, es decir, que por medio del desarrollo de este tipo de proyectos podamos aportar conocimiento para la predicción de sus cultivos del café basándonos en la analítica de datos donde esta se caracteriza por la complejidad de los diversos factores que influyen en los procesos y los beneficios que puede aportar el Big Data con las tecnologías asociadas son cada vez más accesibles.

De igual modo sería una gran entrada para nuestro país y para aquellas personas interesadas en conocer y determinar las variables que pueden afectar el cultivo, cosecha, producción y rendimiento del café.

# Impactos del Proyecto

## Económico

Debido a su impacto e influencia a nivel nacional e internacional contribuye a la estabilidad e integración económica en Colombia siendo la gran entrada de divisas del país posicionado en lo alto por demostrar su eficiencia en la producción del café manteniéndolo al margen a pesar de las diferentes crisis. Para terminar con un manejo óptimo en la contabilidad e ingresos, a la hora de hacer inversiones en la empresa, y además, beneficia de manera directa en la toma de decisiones acertadas en relación con la creación de programas y proyectos futuros planeados por el gobierno nacional y departamental.

## Tecnológico

El sector agrícola está combinando cada vez más tecnologías como la geolocalización, la monitorización del suelo y las condiciones ambientales, la inteligencia artificial, la computación en la nube, el internet de las cosas (IoT), la analítica de datos, para medir con precisión las variaciones de multitud de variables en los campos de cultivo y así mejorar la producción y el rendimiento del café.

Extender la tecnología en un campo rural, permite que la agricultura acceda a una organización ideal y saque la mayor rentabilidad de sus plantaciones al mismo tiempo une dos áreas que por sí solas no tienen mucha relación, sin embargo, generan grandes cambios y ganancias.

Finalmente, el uso de la tecnología en la agricultura permite mejorar la toma de decisiones, además fortalece el proceso de la planificación permitiendo obtener ventajas competitivas, además ayuda al entendimiento de las variables macroeconómicas tales como oferta, demanda, precios de insumos, recursos, maquinaria, costos asociados a la producción y cultivo del café, entre otros.

1. **Social**

Es la forma de integración de pequeños, medianos y grandes caficultores que entrelazan vínculos de amistades o familias entre la sociedad y plasman la cultura cafetera por todos los rincones de Colombia e incluso esto llega a influir mucho a nivel mundial. En general, la sociedad se beneficia al contar con información detallada y conocimiento del proceso del cultivo del café en Colombia.

1. **Ambiental**

Están implementando los cafetales orgánicos y están suprimiendo el uso de insecticidas en la producción del café, optimizando los recursos naturales como el agua, previendo catástrofes naturales, prevención de plagas, etc., gracias al uso del Big data y la tecnología en los procesos agrícolas.

# Restricciones o riesgos asociados y alternativas de solución

## Restricción de Costo

Como conocemos en esta sección se trabaja el presupuesto que se necesita para llevar a cabo el proyecto. Para su desarrollo se creó una tabla de cotización con todos los gastos que se generan, desde trabajadores hasta herramientas necesarias para su finalización.

Tabla 1, Cotización de la materia prima

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gastos | Cantidad | Detalles | Valor Unitario | Valor Total |
| Personal | 5 | Programador  Científico de datos  Diseñador grafico  Ingeniero  Contador | $2.000.000 | $10.000.000 |
| Internet | 50 MB | Fibra Óptica | $100.000 | $100.000 |
| Computador | 5 | **CPU**: Intel Core i5 8900, 2.8 Ghz  **RAM**: 8 GBx2  **Unidad Almacenamiento:** SSD 1 TB | **CPU:** $652.121  **RAM:** $132.000  **Unidad de almacenamiento:**  $345.240  **Total:** $1.129.361  **Con periféricos:** $2.500.000 | $12.500.000 |
| Papelería | 1 | Impresora Laser  Grapadora  Resma de papel  Carpetas  Sobre de manila  Ganchos  Huellero  Saca ganchos  Perforadora | $749.000  $8.000  $7.000  $500  $400  $2.300  $5.000  $6.000  $8.000 |  |
| Energía eléctrica | 1 | Bifásica  240 V | $89.000 | $89.000 |
| Bases de datos | 2 | MySQL  Excel |  |  |
| Licencias de software | 4 | Anaconda  RStudio  Power Bi  Office 365 | Gratis  Gratis  Gratis  $250.000 Familiar | $250.000 |
| Total |  |  |  | $22.489.000 |

## Restricción de Tiempo

Respecto a la duración del proyecto, se designó un cronograma que duraría aproximadamente tres meses (como se observa en las tablas 2, 3 y 4). En los cuales se hicieron aportes claves para la terminación de este.

Tabla 2, Notas del mes Septiembre

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SEPTIEMBRE | | | | | | |
| D | **L** | **M** | **Mi** | **J** | **V** | **S** |
| 30 | **31** | **1** | **2** | **3** | **4** | **5** |
| 6 | **7** | **8** | **9**  **Búsqueda dataframe** | **10** | **11** | **12** |
| 13 | **14** | **15** | **16** | **17** | **18** | **19** |
| 20 | **21** | **22** | **23** | **24** | **25** | **26** |
| 27 | **28**  **Práctica de varianza** | **29**  **Informe dataframe** | **30** | **1** | **2** | **3** |

Tabla 3, Notas del mes Octubre

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| OCTUBRE | | | | | | |
| D | **L** | **M** | **Mi** | **J** | **V** | **S** |
| 27 | **28** | **29** | **30** | **1** | **2** | **3** |
| 4 | **5** | **6** | **7** | **8** | **9** | **10**  **Modelo predictivo Jupyter** |
| 11 | **12** | **13** | **14** | **15** | **16**  **Entrega informe** | **17** |
| 18 | **19** | **20** | **21** | **22** | **23** | **24** |
| 25 | **26** | **27** | **28**  **Diapositivas informe** | **29** | **30** | **31** |

Tabla 4, Notas del mes Noviembre

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| NOVIEMBRE | | | | | | |
| D | **L** | **M** | **Mi** | **J** | **V** | **S** |
| 1 | **2** | **3** | **4** | **5** | **6** | **7** |
| 8 | **9**  **Diapositivas proyecto final** | **10** | **11** | **12** | **13** | **14** |
| 15 | **16** | **17**  **Práctica lenguaje R** | **18** | **19** | **20** | **21** |
| 22 | **23** | **24** | **25** | **26**  **Instrucciones Rstudio** | **27** | **28** |
| 29 | **30** | **1** | **2** | **3** | **4** | **5** |

## Restricción de Alcance

* Predisposición económica, esta es necesaria para cubrir con inconvenientes a lo largo del proyecto, por ejemplo, con la ampliación del cronograma por algún imprevisto..
* Agregar personal calificado para cumplir de manera acertada con los calendarios y cronogramas de actividades.
* La validación del alcance será responsabilidad del equipo del proyecto, quienes estarán monitoreando continuamente los avances del trabajo, y sobre todo el desarrollo oportuno de los entregables.

Hay múltiples riesgos los cuales se deben tener en cuenta como las plagas, las guerrillas, problemas de mercado y clima. Todo esto se debe a que tanto el gobierno como parte de los campesinos tienen fallas debido a que algunos ciudadanos se sienten abandonados por el gobierno o la falta de motivación al no tener producción viable para sus cosechas y por otro lado la falta de administración que el gobierno ejerce en apoyo para ellos.

Para solucionar estos problemas de producción y hacer que el problema de los campos de café que estén mal administrados por parte de los campesinos será establecer nuevamente el trazado de tierra para establecer la distancia entre cada planta de café, para evitar deslizamientos del lote. Gracias al análisis de dato de Big Data podremos hacer que este sea mucho más productivo que antes y podremos hacer que el gobierno este más interesado en ellos por la rentabilidad que se les está dando y tener más impacto en la economía del país.

# Productos o resultados del desarrollo del proyecto

## Metodología Empleada

***Metodología CRISP-DM - Cross Industry Standard Process for Data Mining.***

**CRISP-DM** (del inglés Cross Industry Standard Process for Data Mining),​ traducido como **Proceso estándar de la industria para la minería de datos,** se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos. Es el modelo analítico más usado. Metodología cubre de forma normalizada el ciclo de manera secuencial, como se haría un proyecto de desarrollo de software.

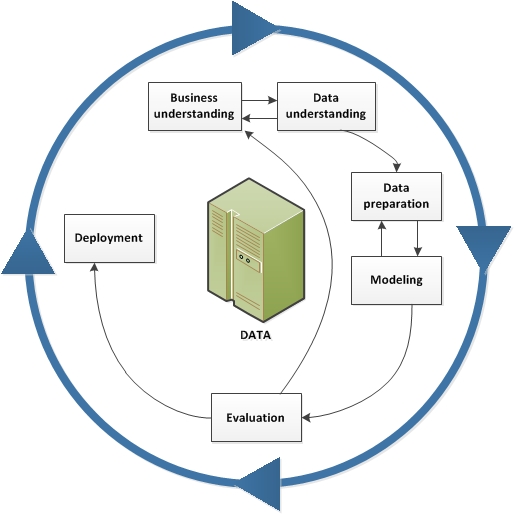


Ilustración 3, Metodología Crips-Dm, Fuente: https://www.ibm.com/support/knowledgecenter/es/SS3RA7\_sub/modeler\_crispdm\_ddita/clementine/crisp\_help/crisp\_overview.html (último acceso: 26 de 01 de 2021).

CRISP-DM fue concebido en 1996. En 1997 se puso en marcha como un proyecto de la Unión Europea bajo la iniciativa de financiación ESPRIT. El proyecto fue dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR y Ohra, una compañía de seguros. Este modelo cubre las fases del proyecto, sus respectivas relaciones entre tareas, además de poder avanzar o retroceder entre fases si es necesario por lo que esto hace al modelo algo flexible.

En esta sección se describe el lenguaje de programación, las aplicaciones de interfaz de programación, las librerías y el software para la realización de este trabajo.

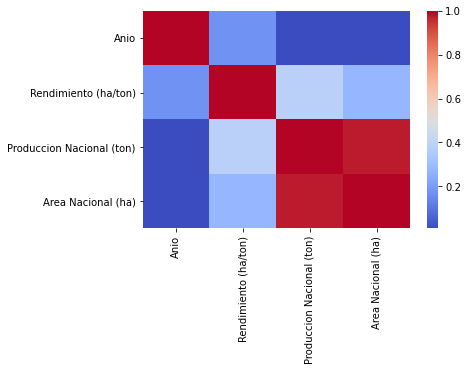
* **Jupyter Notebook:** “Software de código abierto, estándares abiertos y servicios para computación interactiva”, la cual permitió una correcta inspección de los datos. Por otro lado, su interfaz gráfica atribuyo al análisis de escenarios que se encontraban relacionados, entre ellos:
* Pandas Profiling: realiza un análisis exploratorio de datos, y genera informes interactivos en formato web que pueden ser presentados a cualquier persona
* Numpy: crear vectores y matrices grandes multidimensionales
* Seaborn: permite generar fácilmente elegantes gráficos, como mapas de calor que se representa en la siguiente figura.

Ilustración 4, Mapa de calor que relaciona las variables utilizando la librería Seaborn

* **RStudio:** Entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Con su ayuda se obtuvo un amplio conocimiento de Paquetes e instrucciones capaces de resaltar y graficar las estadísticas, por ejemplo:
* Matplotlib: genera gráficas a partir de datos contenidos en listas, vectores
* Dplyr: diseñada para permitir la manipulación de marcos de datos
* Ggplot2: devuelve subconjuntos de vectores, matrices o marcos de datos que cumplen las condiciones.
* Readr: proporciona una forma rápida y sencilla de leer datos rectangulares como csv, tsv y fwf
* Gapminder: delimitada por tabuladores, un conjunto de datos sin filtrar más grande, esquemas de color prefabricados para los países y continentes
* Psych: es una caja de herramientas de propósito general para la personalidad, la teoría psicométrica y la psicología experimenta, como la siguiente gráfica realizada con la instrucción: **pairs.panels(cafe\_df[c(4,5,6)])**

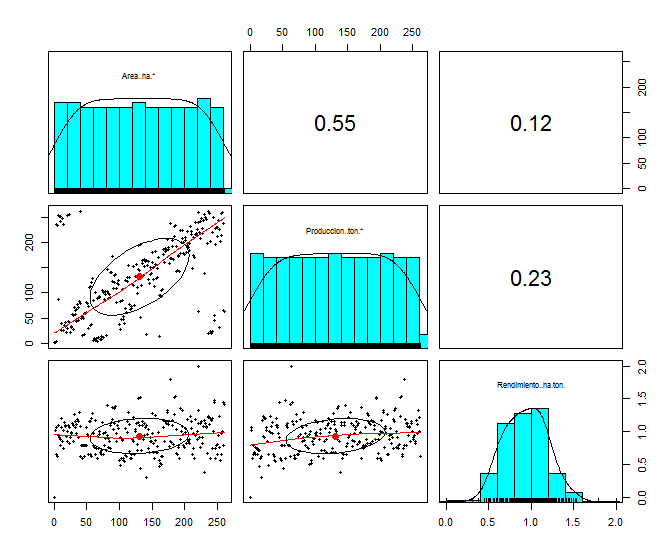
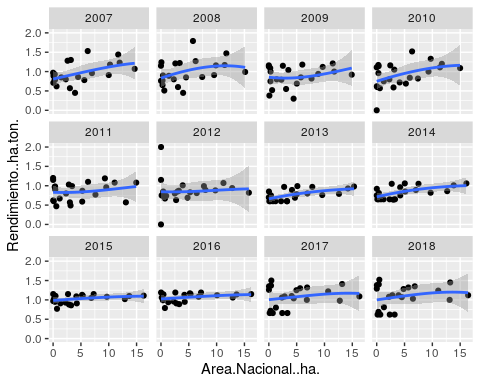
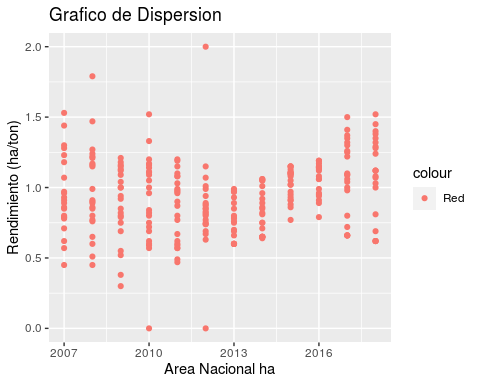


Ilustración 5, Histogramas y polígonos entre el rendimiento, área y producción nacional

Posteriormente veremos algunas graficas que pueden ser creadas con la ayuda de la variedad de librerías que nos ofrece RStudio, además de que hacen mas llamativas la personalización de las tablas.



**Ilustración 6, Dispersión del Rendimiento del café acumulado por año**

Ilustración , Rendimiento por Area nacional en secuencia de los años

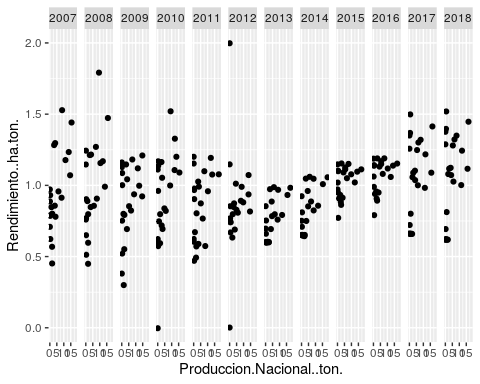
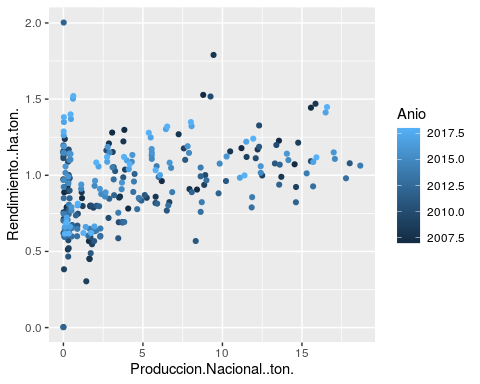
 

Ilustración 9, Correlación entre la Producción vs Rendimiento del café

Ilustración 8, Dispersión en paneles por año

## Requerimientos de los datos

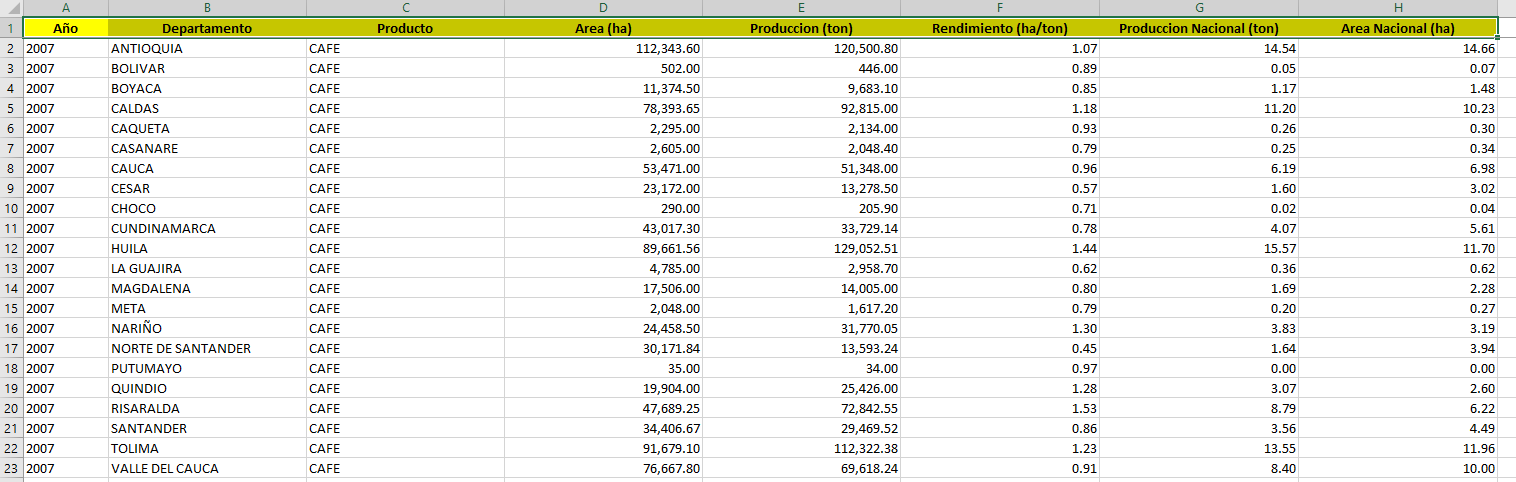
*  Dataframe tipo semi-estructurado, debido a que se encuentra guardado en una hoja de cálculo con formato CSV (Separado por comas).

Ilustración 10, Estructura de los datos utilizados en el análisis, Fuente: propia

* Datos sectoriales recopilados a partir de páginas fiables como la federación nacional de café Colombiano (FNC), la FAOSTAT, Agronet, entre otros.
* Información sobre las variables del café, tales como año, departamento, Área (ha), Producción (ton), Rendimiento (ha/ton), Producción Nacional (ton), Área Nacional (ha).

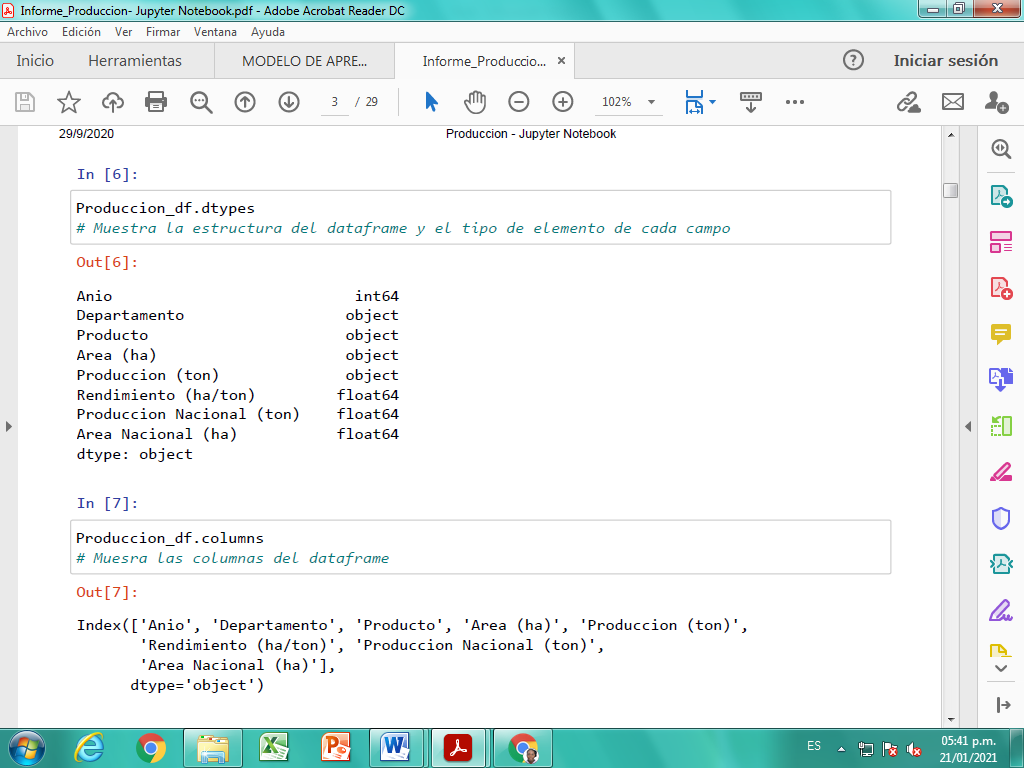


Ilustración 11, Variables del dataframe café, Fuente: propia

El siguiente reporte dispone de atributos más exactos del Rendimiento del café, donde se amplía las descripciones numéricas de dicha variable, dado su formato interactivo se pueden obtener los resultados deseados de todas las características recogidas del café Colombiano.

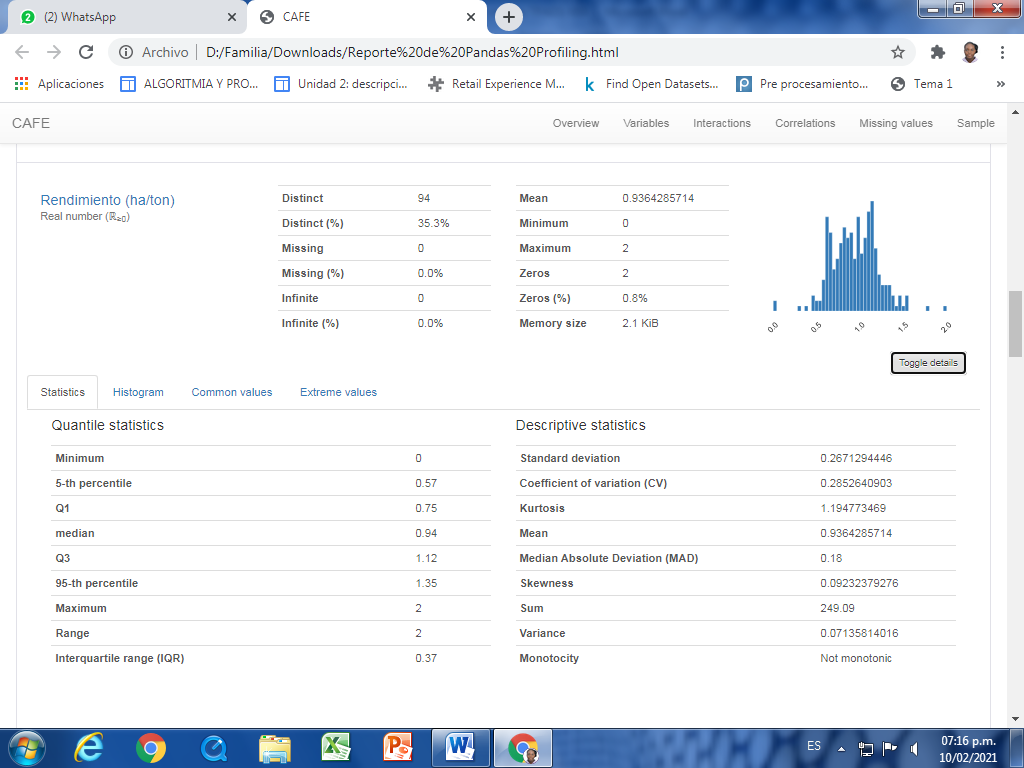


Ilustración 12, Toma de reporte interactivo de Pandas profiling

## Entendimiento de los datos

**Herramientas** **Informáticas**: Rstudio, Jupyter notebook, Excel, Pandas Profiling.

**Técnicas**: Correlación, Desviación Estándar, Media, Percentiles, etc.

Los siguientes resultados que se presentan a continuación, fueron elaborados utilizando las anteriores herramientas y técnicas, para lo cual se tiene alojado todos estos informes detallados en los repositorios de GITHUB y Google Drive encontrados al final.

Tabla 5, Diagnostico estadístico de la mensual del café Colombiano

|  |  |
| --- | --- |
| *Diagnostico estadístico de la producción mensual del café Colombiano desde 1956 - 2020* | |
|  |  |
| Media | 875,591495 |
| Error típico | 13,7358744 |
| Mediana | 821,5 |
| Moda | 644 |
| Desviación estándar | 382,637184 |
| Varianza de la muestra | 146411,214 |
| Curtosis | 1,25212465 |
| Coeficiente de asimetría | 0,85832437 |
| Rango | 2628 |
| Mínimo | 127 |
| Máximo | 2755 |
| Suma | 679459 |
| Cuenta | 776 |

**Test de Shapiro Wilk:** el diagnostico que nos arroja el modelo empleado indica que el valor entregado en la variable es MENOR a **0.05**, nos muestra que si existe una distribución normal y correlación entre las variables, así mismo se observa en las gráficas de dispersión que los puntos mientras más dispersos estén de la línea menos correlación tendrá una variable de otra.

*# importar la Liberia shapiro para realizar el TEST DE SHAPIRO WILK,*

*# el test de Shapiro Wilk CONFIRMA EFECTIVAMENTE la correlación entre las variables* **from** scipy.stats **import** shapiro

estadistico,p\_value **=**shapiro(produccion\_df['Rendimiento (ha/ton)'])

print('Estadística=%.3f, El Valor de: p\_value=%.3f' **%** (estadistico,p\_value))

*# Si el valor entregado en la variable P\_VALUE es MENOR a 0.05,*

*# indica que si existe una distribucion normal y correlacion entre las variables*

Estadística=0.983, El Valor de: p\_value=0.003

**Modelo de Pearson:**

*# valores correlacion de Pearson*

**import** numpy **as** np

produccion\_correlacion\_pearson **=** produccion\_df.corr(method**=**'pearson') produccion\_correlacion\_pearson

*# mide la relación estadística entre dos variables continuas. ... El coeficiente de correlación puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva.*

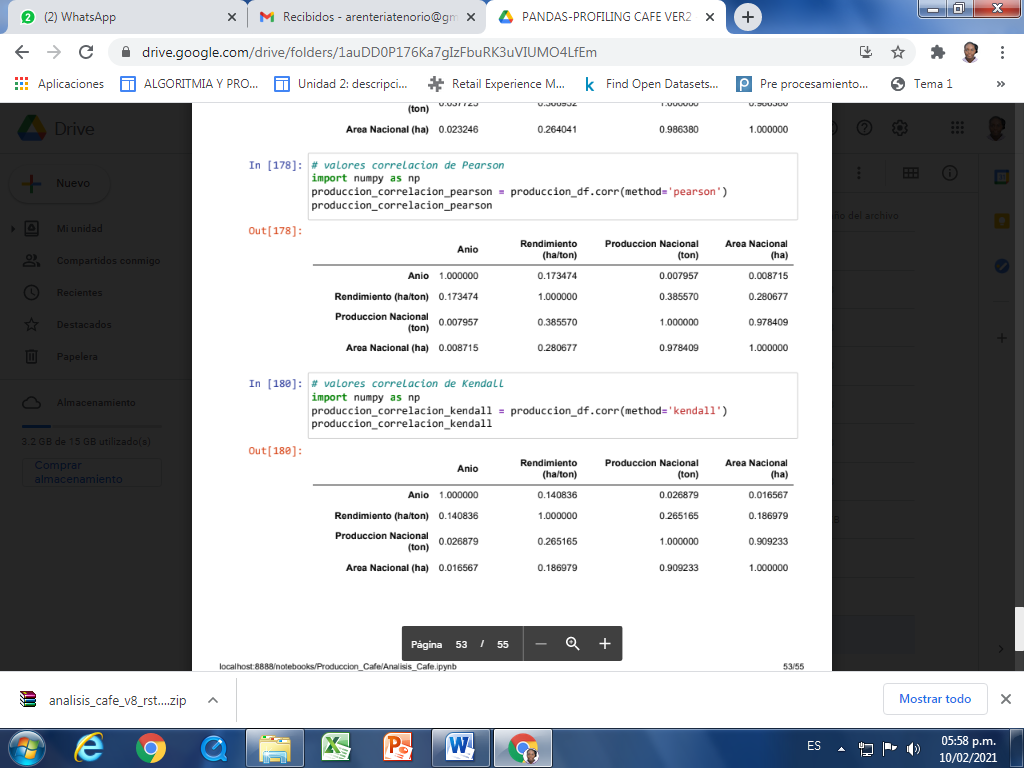


Ilustración 13, Modelo de Pearson

**Modelo de Spearman:**

*# valores correlacion de Spearman*

**import** numpy **as** np

produccion\_correlacion\_spearman **=** produccion\_df.corr(method**=**'spearman') produccion\_correlacion\_spearman

*# Los valores del COEFICIENTE DE SPEARMAN cercanos a cero o inferiores a (+-)(0.*

*# indica que las variables no tienen correlación*

*# Los valores del COEFICIENTE DE SPEARMAN mayores a (+-)(0.4) o indica que las v*

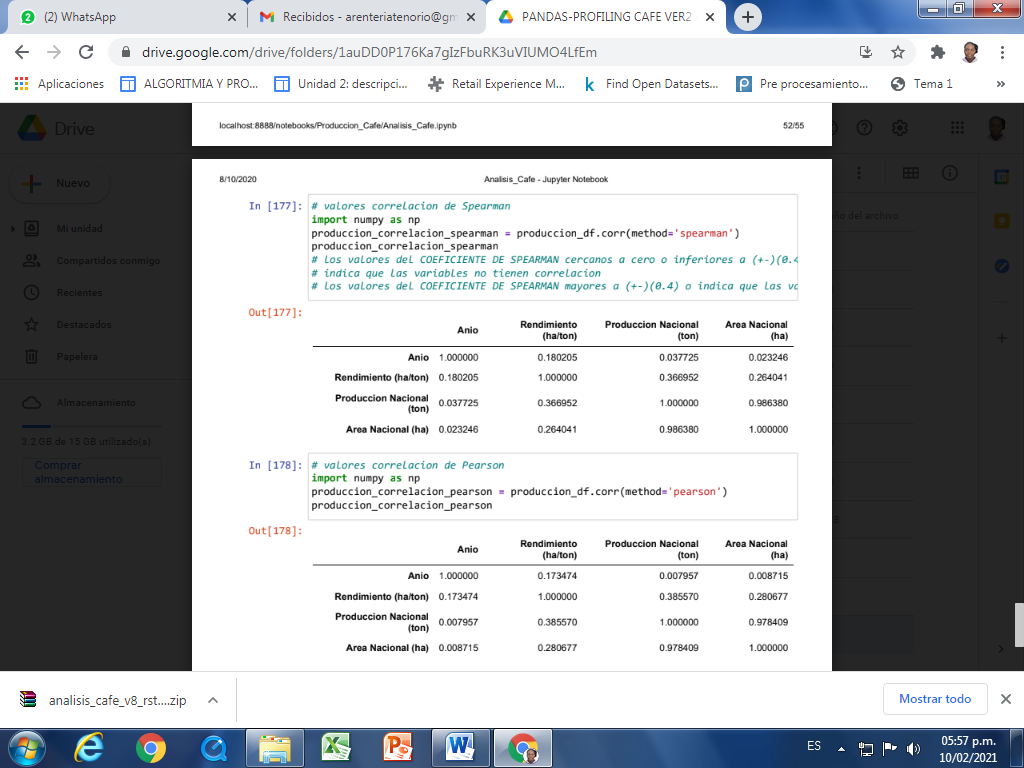


Ilustración 14, Modelo Spearman

**Modelo de Kendall:**

*# valores correlacion de Kendall*

**import** numpy **as** np

produccion\_correlacion\_kendall **=** produccion\_df.corr(method**=**'kendall') produccion\_correlacion\_kendall

*# Mide el grado de asociación entre varios conjuntos (k) de N entidades. Es útil para determinar el grado de acuerdo entre varios jueces, o la asociación entre tres o más variables.*

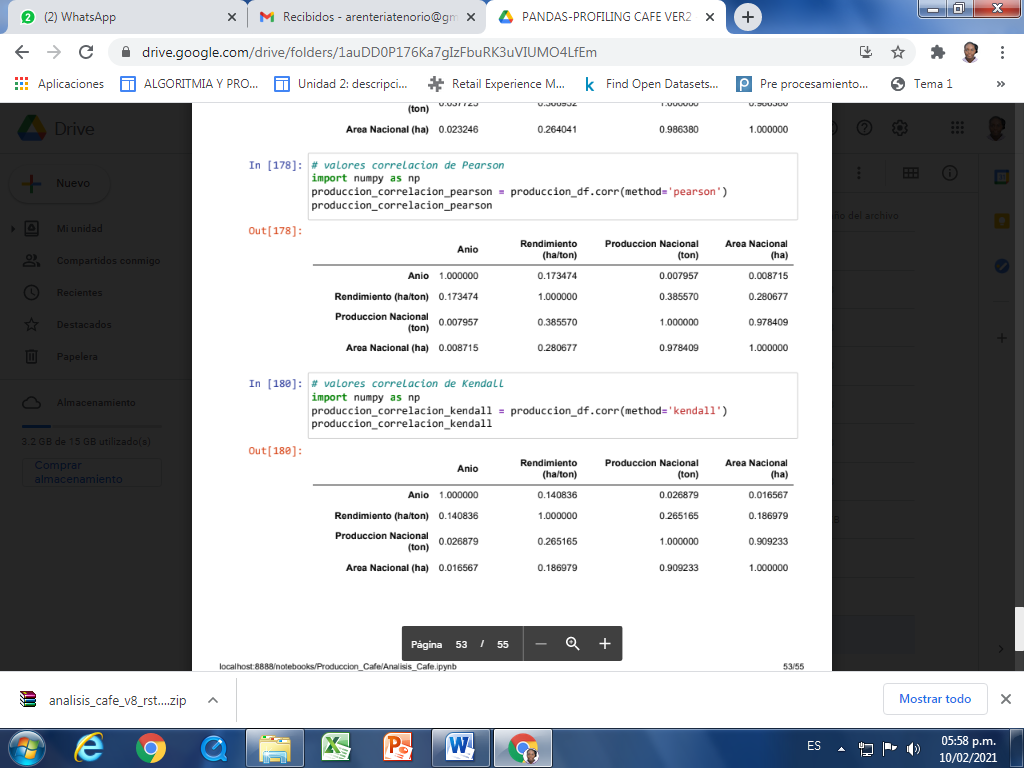


Ilustración 15, Modelo Kendall

## Preparación de los datos

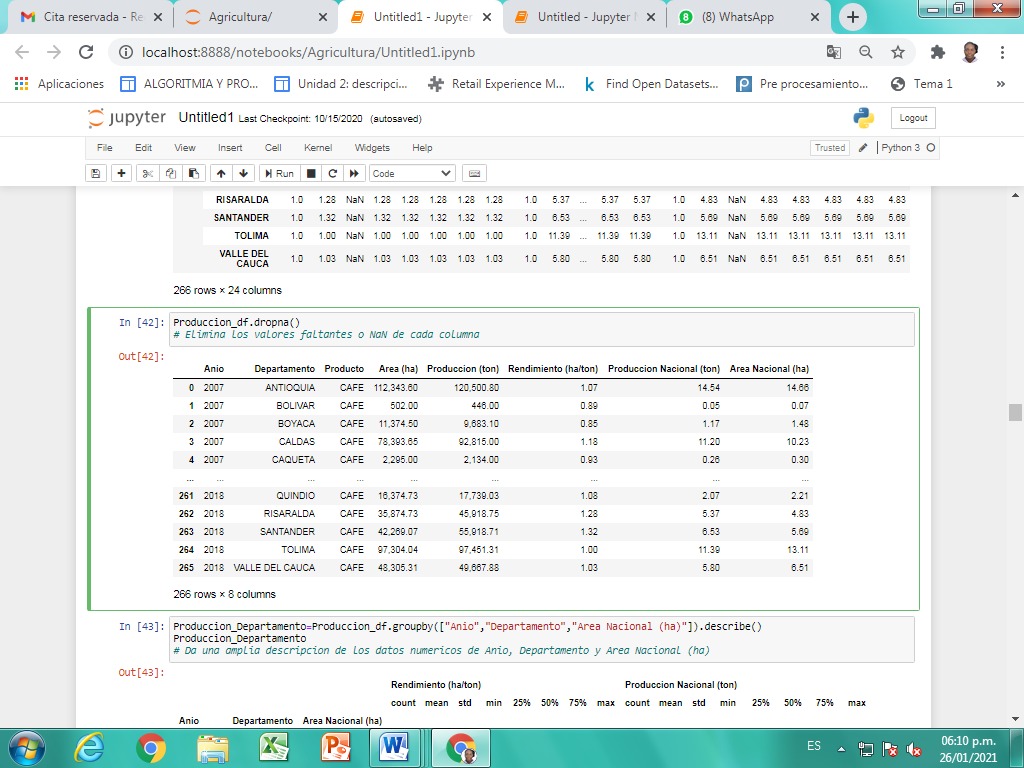
Para la limpieza y transformación de los datos se emplearon una serie de pasos, implementando algoritmos y aplicaciones Big data para que estos se conviertan en datos útiles, limpios y de importancia para tener buenos resultados en su análisis.

Proceso ETL, este acrónimo representa sus siglas en ingles que son:

* **Extract:** Se toma la información de su origen
* **Transform:** Se adecuan los datos extraídos y se convierten en datos óptimos para su utilización
* **Load:** Se cargan los datos que fueron transformados

A continuación veremos algunas instrucciones que facilitaron la búsqueda y limpieza de OUTLIERS en nuestro dataframe.

* **Dropna ():** elimina valores faltantes



**Ilustración 16, Eliminación de valores faltantes en Python**

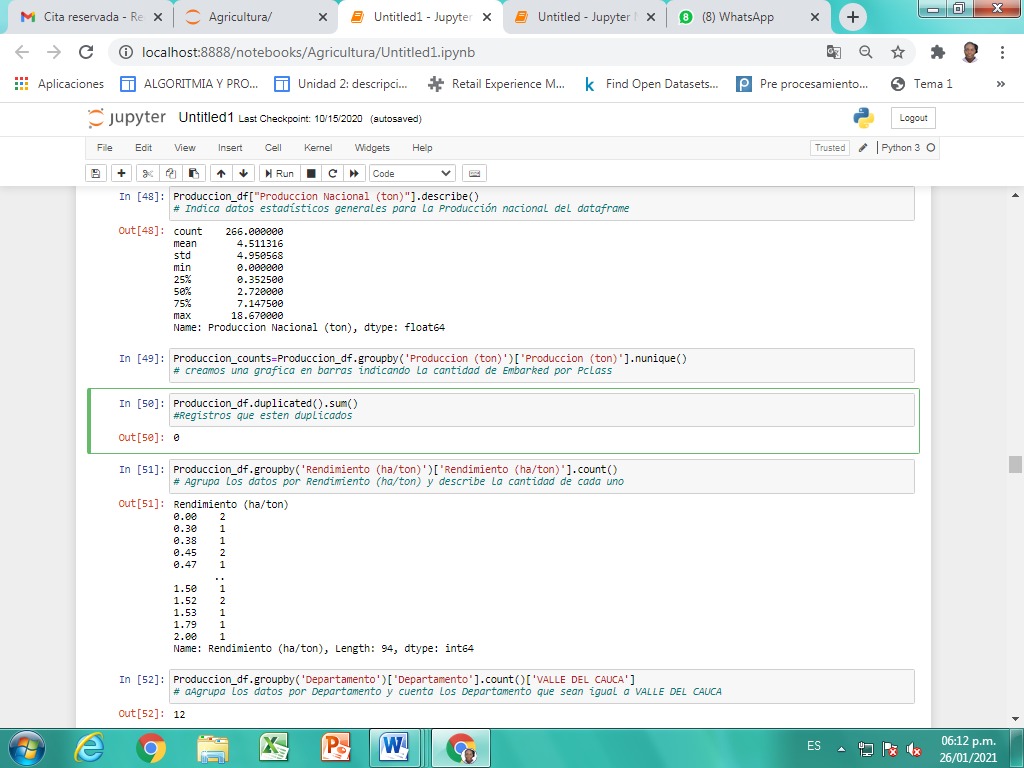
* **Duplicated ().sum ():** suma los registros que se encuentren duplicados

Ilustración 17, Conteo de valores duplicados en Python

* **Is.na(Cafe\_df$Area..ha.):** verifica si existen valores nulos

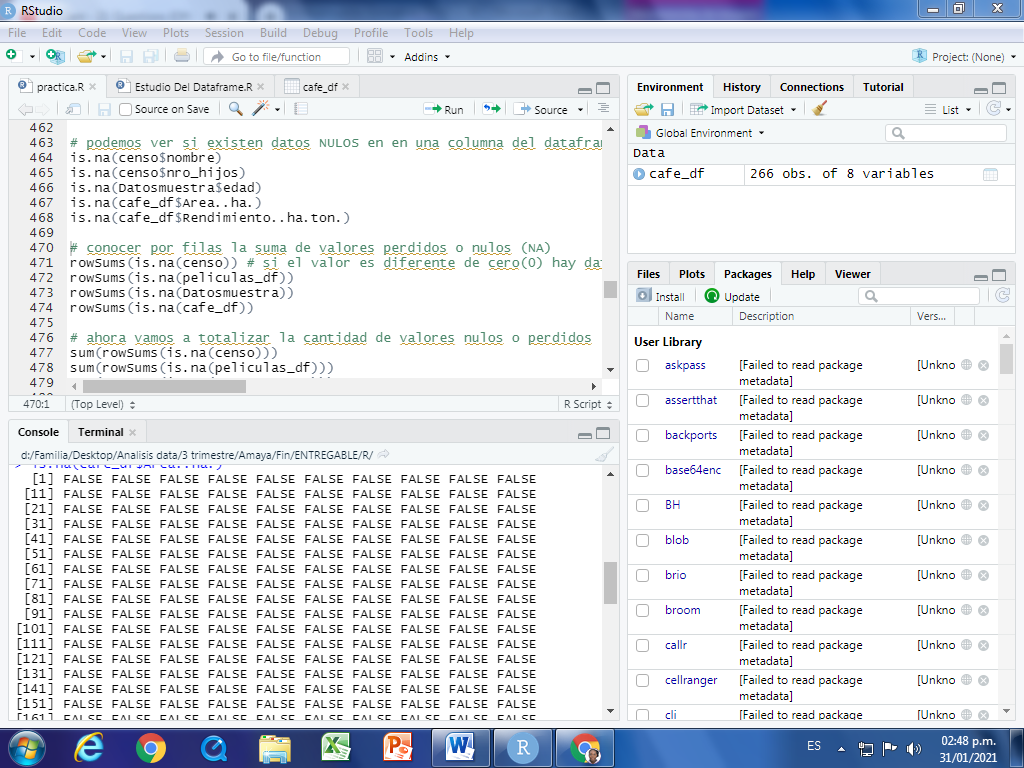


Ilustración 18, Inspección de valores nulos en R

* **RowSums(is.na(cafe\_df)):** conocer por filas la suma de valores perdidos o nulos (NA), si el valor es diferente de cero(0) hay datos nulos

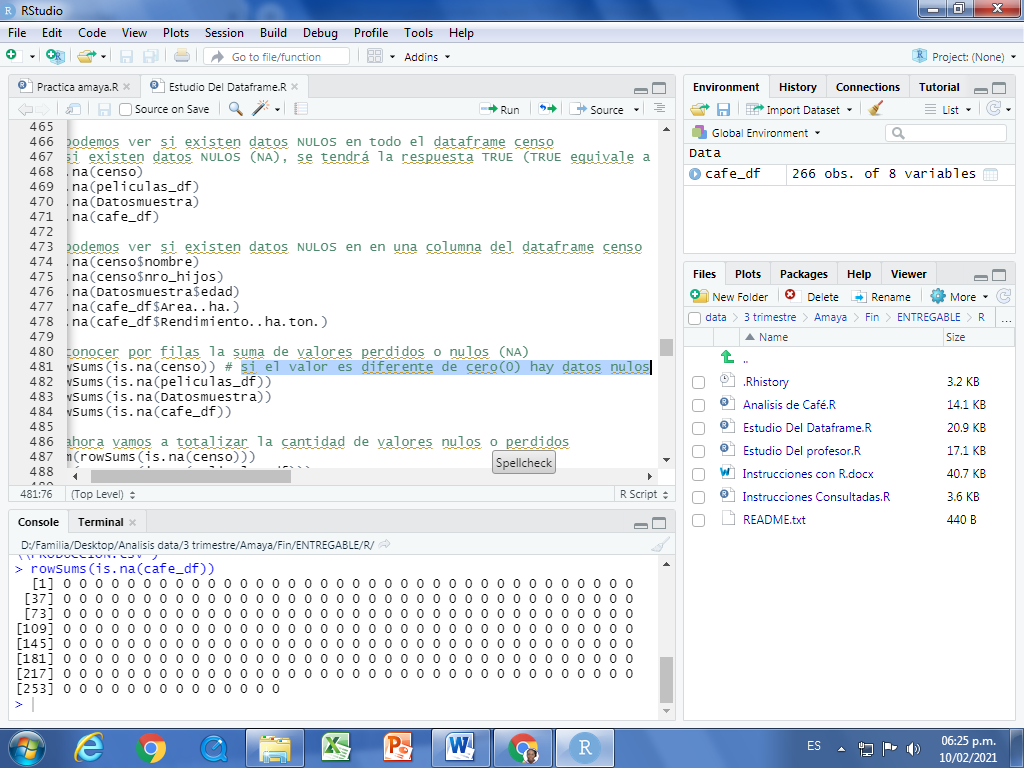


Ilustración 19, Columna de valores perdidos

## Preparación del modelo

Modelo de regresión lineal, La regresión lineal permite modelar la relación entre unas variables de entrada (explicativas) con valores de salida reales. la siguiente ecuación, válida la instrucción para recrear el modelo.

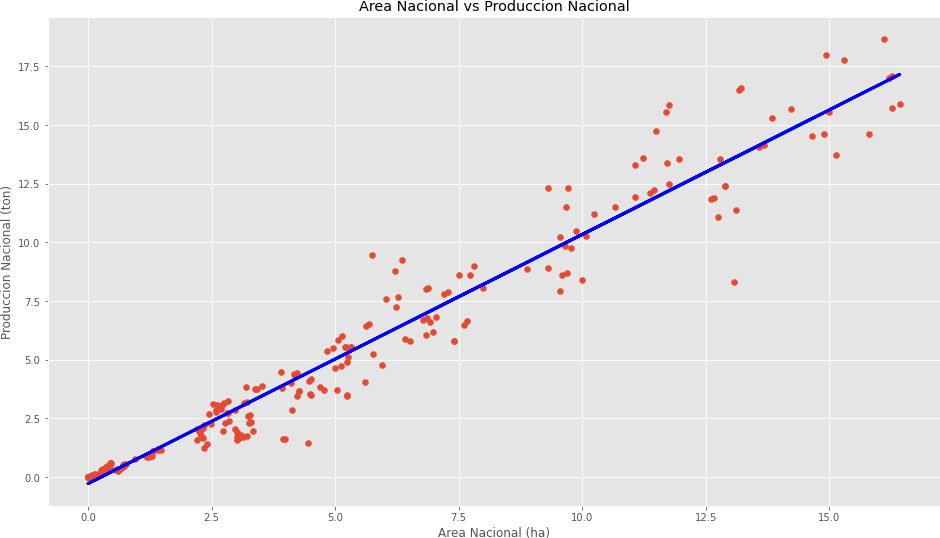


Ilustración 20, Visualización del modelo lineal en Python

Llamada **ecuación canónica**, en donde:

**Y= mX+bo**

* M: es la pendiente de la recta
* Bo; es el intercepto con el eje Y

Por otro lado, en el caso de R el modelo de regresión lineal se crea con la siguiente expresión:

**Grafica\_AreaNal\_Rendim\_1=ggplot(cafe\_df, aes(x = Area.Nacional..ha., y = Rendimiento..ha.ton., color ="red")) +**

**geom\_point()**

**Grafica\_AreaNal\_Rendim\_1**

# Genera un gráfico de dispersión entre las variables Área Nacional vs Rendimiento de color rojo

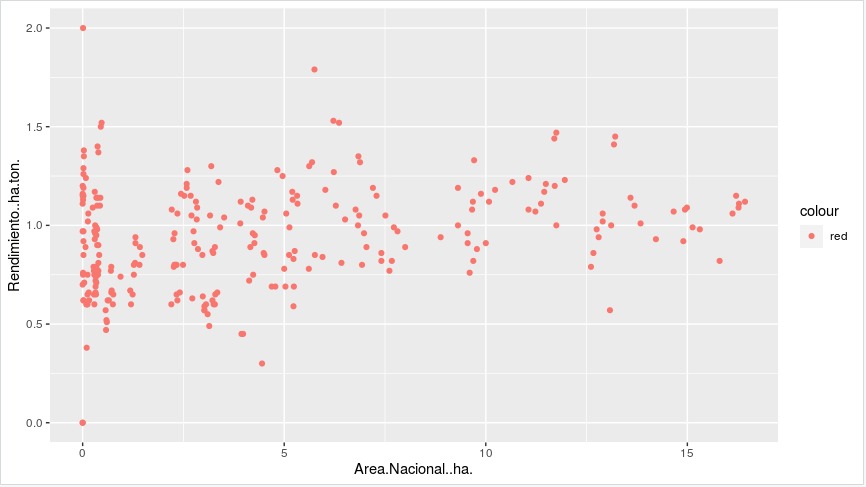
****

Ilustración 21, Grafico de dispersión Área N vs Rendimiento en RStudi

**Grafica\_AreaNal\_Rendim\_1 + geom\_point()+geom\_smooth(method="lm", colour ="Blue")**

# Define los ejes X & Y y traza un modelo lineal que se aleje lo menos posible de los datos

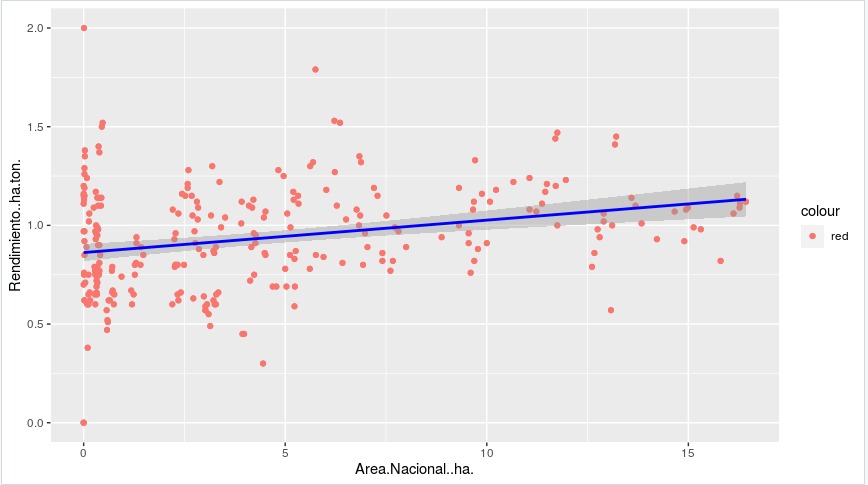


Ilustración 22, Modelo de regresión lineal Área n. vs Rendimiento en RStudio

## Evaluación del Modelo

Las pruebas hechas en el dataframe “PRODUCCIÓN” fueron realizadas a través de la instrucción **‘Produccion\_Obtenida = Regr\_predict ([[X]])**, la cual permite hacer una estimación o predicción de una variable a futuro

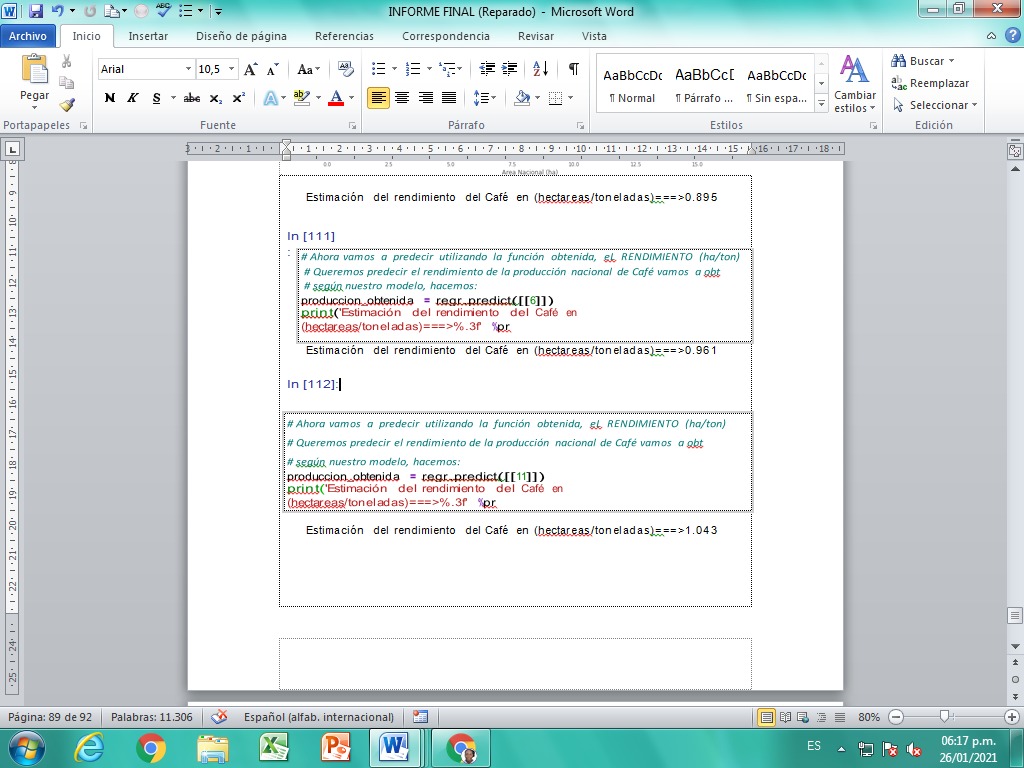
****

Ilustración 23, Grafica estimación del Área Nacional cosechada por el Rendimiento del café

# Entregables y su descripción

Para facilitar el acceso a los archivos que se van a entregar como evidencia del desarrollo del proyecto formativo, se creó una carpeta en **GitHub** enlace: <https://github.com/robinjoot/proyecto-formativo-.git> donde se encontrarán todos los documentos realizados a lo largo del proyecto y en el **Google Drive** disponible en: <https://drive.google.com/drive/folders/1JwA2oeB9n7xlt4O457IWNo1z93U3f7tL?usp=sharing>

# Conclusiones

Con el desarrollo de este proyecto de investigación, es posible obtener el mejor modelo de predicción, que pueda identificar y clasificar variables que afectan directamente la producción del cultivo de café en Colombia otorga a la ciencia de los datos incrementar las probabilidades de éxito en el sector agrícola y otros sectores que comienzan a gestionar su transformación en sus procesos, productos o servicios para adaptarse a la dinámica de los mercado actuales implementando técnicas Big Data que les permita tomar decisiones más inteligentes y fortalecer todos los procesos (producción, humana, tecnológicos, manejo de datos, mercadeo, etc.), que les permite ser más competitivas.

De esta forma, deben hacerse de tecnologías BIG DATA asociadas en la generación, captura, procesamiento y almacenamiento de sus propios datos de forma dinámica. Finalmente con la ayuda de los modelos empleados en el transcurso, se lograron analizar y estructurar los datos y observar los resultados obtenidos para determinar sí las soluciones son o no las indicadas, lo que permitirá predecir comportamientos importantes en el cultivo y la producción del café a nivel nacional y permitiría su exportación a una gran cantidad de países, con mejores divisas.

# Bibliografía

Elenco, H. (s.f.). *AgroInteligencia*. Obtenido de https://www.agrointeligencia.com/big-data-agricultura-enrique-serrano/

*Federacion Nacional de Cafeteros*. (s.f.). Recuperado el 28 de 01 de 2021, de https://federaciondecafeteros.org/wp/servicios-al-caficultor/aprenda-a-vender-su-cafe/

González, A. (2019). *CleverData*. Obtenido de https://cleverdata.io/que-es-machine-learning-big-data/#:~:text=Machine%20Learning%20es%20una%20disciplina,crea%20sistemas%20que%20aprenden%20autom%C3%A1ticamente.&text=Autom%C3%A1ticamente%2C%20tambi%C3%A9n%20en%20este%20contexto,el%20tiempo%2C%20sin%20i

*IBM*. (s.f.). Recuperado el 26 de 01 de 2021, de https://www.ibm.com/support/knowledgecenter/es/SS3RA7\_sub/modeler\_crispdm\_ddita/clementine/crisp\_help/crisp\_overview.html

Intel. (s.f.). *Intel*. Obtenido de https://ark.intel.com/content/www/es/es/ark/products/126687/intel-core-i5-8400-processor-9m-cache-up-to-4-00-ghz.html

(s.f.). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* Informatico.

Pastor, I. D. (24 de 9 de 2019). BIG DATA en la agricultura. (P. Bolivar, Entrevistador)

Peña, J. A. (2019). *Proyecto de grado.* Investigativo, Bogota, Colombia.

Ramírez, D. H. (2018). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD.* Pereira.

# Recomendaciones

* Resulta importante que en todo proceso de Analítica de datos se analicen, procese, recolecte, filtre, depure muy bien los datos para finalmente conseguir una óptima ejecución
* En este proceso del Big Data se utilizó una metodología llamada CRIPS-DM, debido a que es la más adecuada ya que orienta todas las actividades durante el tratamiento de datos
* Somos el primer curso de Programación para Analítica de datos por tal motivo se debe mejorar aspectos y realizar ajustes para las siguientes generaciones
* Es recomendable poder verificar y validar el modelo presentándolo a Instituciones, entidades y Organizaciones que lo prueben en la práctica real.