ANALÍTICA DE DATOS PARA EL RENDIMIENTO EN LOS CULTIVOS DE AGUACATE HASS EN COLOMBIA

Cristian Augusto Lozano Vásquez

Juan Esteban Suaterna Cabrera

Universidad Externado de Colombia

Facultad de Administración de Empresas

Maestría Gerencia Estratégica en Tecnologías de la Información

Bogotá, D.C. – Colombia

2019

ANALÍTICA DE DATOS PARA EL RENDIMIENTO EN LOS CULTIVOS DE AGUACATE HASS EN COLOMBIA

Proyecto de Investigación

Cristian Augusto Lozano Vásquez

Juan Esteban Suaterna Cabrera

Jorge Mario Calvo Londoño

Asesor

Universidad Externado de Colombia

Facultad de Administración de Empresas

Maestría Gerencia Estratégica en Tecnologías de la Información

Bogotá, D.C. – Colombia

2019

Agradecimientos

Agradecemos a cada uno de los profesores que compartieron su conocimiento y amor por el estudio, en especial al Profesor Jorge Mario Calvo, por la confianza y por ser el mentor del conocimiento adquirido durante el desarrollo de este proyecto, por último, a la Universidad Externado de Colombia por permitirnos ser parte de la comunidad externadita.

Agradezco profundamente a Dios por ser mi guía, mi amigo y bendición en cada paso de la vida, a mi madre Belsy por el amor y sacrificio innato de su ser para sus hijos, a mi padre Daniel (Q.E.P.D) por su amor y apoyo desde el cielo, a mi abuela Rosalba (Q.E.P.D) por enseñarme el verdadero amor incondicional, a mi abuelo Germán por inculcarme el valor de los principios y la perseverancia, a mi tío Mario por enseñarnos el valor de la familia y el apoyo incondicional. Agradezco a mi esposa Estella por su amor y comprensión, a mi hijo Cristian Daniel por ser mi alegría, inspiración y fortaleza.

Cristian Augusto

Agradezco a Dios por permitirme y regalarme esta experiencia tan maravillosa de aprendizaje y de crecimiento personal , profesional, y espiritual porque sin él este logro no habría sido posible, como también agradezco a mi esposa Guiselle por el amor y el apoyo incondicional que me brindó durante el desarrollo de mis estudios, a mis hijas María Paula y Laura Sofía por ser el motor que impulsa mi vida para crecer y superarme cada vez más en todos los aspectos de mi vida; a mis Padres Rafael y Esperanza por enseñarme la fe, convicción y perseverancia que hay que tener para afrontar las diferentes situaciones en la vida.

Juan Esteban

Contenido

Lista de Ilustración

Lista de Tablas

Lista de Gráficas

INTRODUCCIÓN

Existe una tendencia en las empresas, donde las decisiones de gestión dependen menos del "instinto" de un líder y, en cambio, de las analíticas basadas en datos (Brynjolfsson, Hitt, & Kim, 2011). Independiente del sector económico, las empresas buscan ser más competitivas para crecer y perdurar en el tiempo. Por lo tanto, son impulsadas a crear estrategias que puedan consolidarse en una ventaja competitiva con respecto a la competencia en un mercado globalizado, caracterizado por ser reñido y dinámico.

Las empresas que marcan la diferencia son aquellas que gestionan de otra manera, partiendo de reconocer la información como un activo y de la capacidad de transformarse para que los datos y modelos de análisis conduzcan a la generación de conocimiento empresarial y a una mejor toma de decisiones (Rodriguez, 2016).

En el sector agrícola es importante que comience a prosperar la gestión de estrategias basadas en información, teniendo presente la gran incertidumbre de los factores que influyen en la producción de cultivos y en la enorme cantidad de datos que son y pueden ser capturados durante todo el proceso de la producción.

Adicionalmente, son cada vez más las entidades y empresas que hacen accesibles las fuentes de datos centralizadas y las herramientas para la generación de modelos de análisis de datos que den respuestas a las problemáticas particulares y del sector.

Actualmente, Colombia tiene la oportunidad de convertirse en unos de los principales exportadores de aguacate Hass. Sin embargo, existe una problemática en cuanto al índice de rendimiento (expresado en toneladas por hectárea) de su producción en comparación con otros países exportadores porque desconoce los factores que influyen en la producción del cultivo.

Teniendo en cuenta lo anterior, se realiza el planteamiento del presente proyecto de investigación, con el objetivo de dar respuesta a la problemática por medio de la generación de un modelo de analítica de datos basado en una metodología de minería de datos y haciendo uso de una técnica de aprendizaje de máquina supervisada que, a diferencia de la estadística clásica, centra en gran parte sus objetivos en la predicción de resultados (Crane-Droesch, 2018).

CAPÍTULO 1. PLANTEAMIENTO

1.1 Antecedentes

Con la evolución de las tecnologías de la información han permitido ampliar la visión de las organizaciones para poder entender las necesidades y diferentes comportamientos del mercado, y así ser sostenibles en el tiempo. El sector agrícola no puede ser la excepción en comenzar adoptar las mejores prácticas y herramientas que le permitan al agricultor tomar decisiones e incentivar la inversión agropecuaria y el crecimiento de la economía en el sector. Dicho lo anterior y teniendo en cuenta que diferentes entidades públicas y privadas han venido recolectando diferentes tipos de datos estructurados, semi-estructurados y no estructurados en diferentes escalas, como, por ejemplo, las estaciones meteorológicas, donde producen grandes volúmenes de datos horarios, diarios, mensuales, entre otros, de diversas variables. Razón por la cual, permiten la posibilidad de iniciar un camino de analítica de datos en un sector que tiene la potencialidad de crecer y ser cada vez más rentable sí se incorporan herramientas, metodologías y mejores prácticas de manera preventiva para mitigar riesgos financieros aumentando cada vez más la rentabilidad de los cultivos de Aguacate. Por tal razón es importante articular desde el inicio la analítica de datos con el Big Data.

Es importante hacer énfasis en la toma de decisiones y la estrategia de negocios apoyada en información, particularmente bajo el enfoque empresarial aplicado al sector agrícola. Sector que se caracteriza por la complejidad de los diversos factores que influyen en los procesos y donde los beneficios que puede aportar el Big Data y las tecnologías asociadas son cada vez más accesibles. Por otro lado, la problemática del cambio climático y el impacto en el rendimiento de los cultivos, principalmente en la producción de aguacate Hass en Colombia y la oportunidad económica

asociada, hacen del Big Data y la analítica de datos un tema atractivo y una tendencia para impulsar el desarrollo de investigaciones en el sector agrícola colombiano.

La toma de decisiones y la estrategia de negocios han sido permeadas por el Big Data y la analítica de negocios en casi todos los aspectos (Griffin & Wright, 2015). En el contexto empresarial, el uso de las TI no pueden medirse en términos de rentabilidad o productividad sino en procesos y funciones que tiene impacto en el rendimiento; y así como el valor de la información utilizada de manera efectiva significa una clara ventaja competitiva, el Big Data y las TI agregan valor indirectamente en las organizaciones (Ylijoki & Porras, 2018).

En el sector agrícola, la agricultura de producción depende de muchos factores biológicos, climáticos, económicos y humanos que interactúan de forma compleja. Los productores agrícolas y las empresas de las industrias agrícolas deben tomar innumerables decisiones a diario que impactan en el rendimiento y en la operación de la cadena de suministro respectivamente (Majumdar, Naraseeyappa, & Ankalaki, 2017; Sonka, 2016). Por lo tanto, la toma de decisiones requiere estar mejor fundamentada en diversas fuentes de información que se hacen progresivamente más difíciles de gestionar por fuera del paradigma del Big Data.

La aplicación del Big Data en el sector agrícola, además de poder contribuir optimizando costos y reduciendo el impacto ambiental, identifica como un cambio estratégico la generación de valor con los bajos costos del uso de herramientas de monitoreo y control en actividades agrícolas, la integración de los datos de distintas fuentes para la generación de conocimiento, así como el incremento de la presión para un mejor monitoreo de las actividades agrícolas (Sonka, 2016).

La agricultura es un componente clave en la economía de muchos países, en relación con la seguridad alimenticia, los ecosistemas naturales y el bienestar social (Sonka, 2016). Pero está amenazada por las consecuencias de las intervenciones humanas que han causado estrés en los

ecosistemas naturales y cambios en el sistema climático, agudizando el impacto en el rendimiento de los cultivos por su dependencia con el clima (Crane-Droesch, 2018). La adopción de tecnologías de detección asociadas con Big Data prometen crear valor para los tomadores de decisiones del sector y la sociedad (Sonka, 2016).

Un ejemplo particular de la aplicación Big Data en las fincas está representado por fabricante de maquinaria agrícola John Deere, una empresa exitosa, pionera y a la vanguardia de la innovación que está adoptando técnicas de Big Data con servicios que permiten a los agricultores supervisar en tiempo real los datos recopilados por miles de usuarios con sensores conectados a sus maquinarias mientras trabajan en sus campos y sondas en el suelo, para que puedan tomar decisiones informadas sobre cualquier cosa, desde cultivos a sembrar y fertilizantes a usar. Con la información respecto a las condiciones climáticas, información de las maquinas pueden monitorear y correlacionar los niveles de productividad y reducir los tiempos de inactividad al predecir cuándo y dónde falla un equipo (Marr, 2016).

En el trabajo de investigación titulado “Análisis de datos agrícolas utilizando técnicas de minería de datos: aplicación de Big Data”. Utilizan distintas técnicas de extracción de datos con el objetivo de analizar los datos agrícolas referentes a parámetros de producción y factores climáticos de 5 distintos cultivos (algodón, cacahuete, sorgo, arroz y trigo) en los 28 distritos del estado de Karnataka en la India. Los métodos de evaluación se basaron en los algoritmos de minería de datos, seleccionando el método de agrupación cómo el más óptimo para el análisis de datos agrícolas, describiendo y evaluando los resultados de los métodos de agrupación (PAM, CLARA, DBSCAN), con métricas de calidad. Los métodos facilitaron la identificación de los parámetros óptimos como el clima (temperatura y precipitación) que favorecen la producción de cultivos de

trigo. Además, junto con el método de regresión lineal simple, encontrar los atributos significativos y formar la ecuación para la predicción del rendimiento (Majumdar, Naraseeyappa, & Ankalaki, 2017).

Por otro lado, en el trabajo publicado con el título: "Métodos de aprendizaje automático para la predicción del rendimiento de los cultivos y la evaluación del impacto del cambio climático en la agricultura". Demuestran a través de un enfoque, que utiliza una variante semi-paramétrica de una red neuronal profunda, la relación entre el clima y el rendimiento de los cultivos de maíz, usando datos sobre el rendimiento de maíz del medio oeste de los EE. UU. El enfoque supera a los métodos estadísticos clásicos y las redes neuronales totalmente no paramétricas al predecir el impacto negativo en los rendimientos del maíz usando escenarios de modelos climáticos (Crane-Droesch, 2018).

El sector agrícola en Colombia sigue en crecimiento y se proyecta para ser uno de los motores de la economía, los más favorecidos son los grandes productores y/o gremios organizados. La financiación de cultivos o ubicación de tierras fértiles ya no son inconvenientes críticos, pero persisten los problemas enmarcados en dos grandes componentes: el componente de desarrollo rural, referente a infraestructura, uso eficiente del suelo y propiedad de la tierra; y el componente de productividad, relacionado con el rezago en la productividad y la deficiente comercialización. Se constituyen como factores perjudiciales de la rentabilidad que afectan tanto a pequeños como a grandes productores. Por lo tanto, para lograr un crecimiento rural sostenible y sustentable, adicional al componente de desarrollo rural, es relevante que exista un crecimiento agropecuario enfocado en una mayor productividad, la cual, está vinculada a un gran número de factores que acentúan la incertidumbre sobre el futuro de la inversión en el sector agrícola (Dinero, 2018; Portafolio, 2018; Semana Rural, 2017).

En los últimos años, una gran oportunidad para la producción y exportación de Aguacate Hass se ha venido materializando en Colombia. La demanda mundial del aguacate Hass crece alrededor de un 3%, con un ritmo más acelerado que la producción mundial, abriendo espacio para que otros competidores como Colombia, puedan ocupar el espacio perdido por otros productores y abrirse a nuevos mercados (Dinero, 2017). En la siguiente grafica publicada en el año 2016 por Procolombia, en donde se puede observar la distribución de los países importadores destacándose Estados Unidos, Países Bajos y Francia, entre otros.

Gráfica 1. **Principales importadores de Aguacate 2016**
Fuente: Procolombia, Mercado del Aguacate en Estados Unidos. Obtenido de (PROCOLOMBIA, 2017).

La importancia de realizar la investigación consiste en construir precedentes de investigación en materia de generación de modelos que correlacionen los factores climáticos con el rendimiento de los principales cultivos en Colombia, en nuestro caso con la producción de aguacate Hass, y permitan tomar acciones y generar políticas de adaptación y mitigación mejor sustentadas en base a datos para el sector agropecuario y otros sectores en los que se puedan escalar.

1.2 Planteamiento del Problema

La producción de Aguacate Hass en Colombia, se ha convertido en una gran oportunidad para el país, pero a su vez un gran reto para el agricultor, ya que requiere una gran inversión inicial, pero se ha tratado de materializar a través del conocimiento empírico, es decir, bajo la metodología del ensayo y el error, profundizando el riesgo de perder la siembra y por supuesto la inversión inyectada por los agricultores como consecuencia adversa. Se debe tener en cuenta que hay un dinamismo incremental en el entorno, en donde las manifestaciones de los patrones de comportamiento actualmente no son estables o lineales como lo eran en periodos anteriores. Un ejemplo de ello, es la variación del clima en los últimos tiempos debido al cambio climático del planeta.

Para el análisis estadístico de los cultivos se manejan las siguientes cuatro variables: La primera variable es la siembra que hace alusión al terreno comprometido o destinado a la plantación de la semilla, esta variable se expresa en número de hectáreas (ha), la segunda variable es la Cosecha, que hace alusión al terreno donde florece y genera fruto, esta variable también se expresa en número de hectáreas (ha), la tercera variable es la producción, que hace alusión a la cantidad de fruto generado en un cultivo, esta variable puede estar expresado en kilogramos (Kg) o toneladas (Ton), para nuestro caso vamos hacer uso de la unidad de medida en toneladas, y nuestra cuarta y última variable es el rendimiento, que hace alusión a la cantidad de fruto generado por una hectárea de tierra, esta variable se expresa en (Ton/ha), el análisis de esta última variable va ser nuestro objeto de estudio en el presente trabajo de investigación. De acuerdo a lo anterior se observaron algunas estadísticas de cómo se encuentra Colombia y el departamento de Tolima con respecto a las variables antes mencionadas para entender la oportunidad de analizar la variable

del rendimiento, y así identificar los factores que la impactan tanto positiva como negativamente los cultivos de Aguacate.

Realizando un análisis de las estadísticas presentadas por parte de *Food and Agriculture Organization Corporate Statistical Database* (FAOSTAT), para el año 2016 acerca de las siguientes variables: Cosecha, Producción y Rendimiento, observamos lo siguiente: en términos de cosecha, Colombia se encuentra ubicada en una tercera posición con 35114 hectáreas cosechadas para ese año de estudio, antecedido por México y Perú, como se observa en la gráfica No. 2.

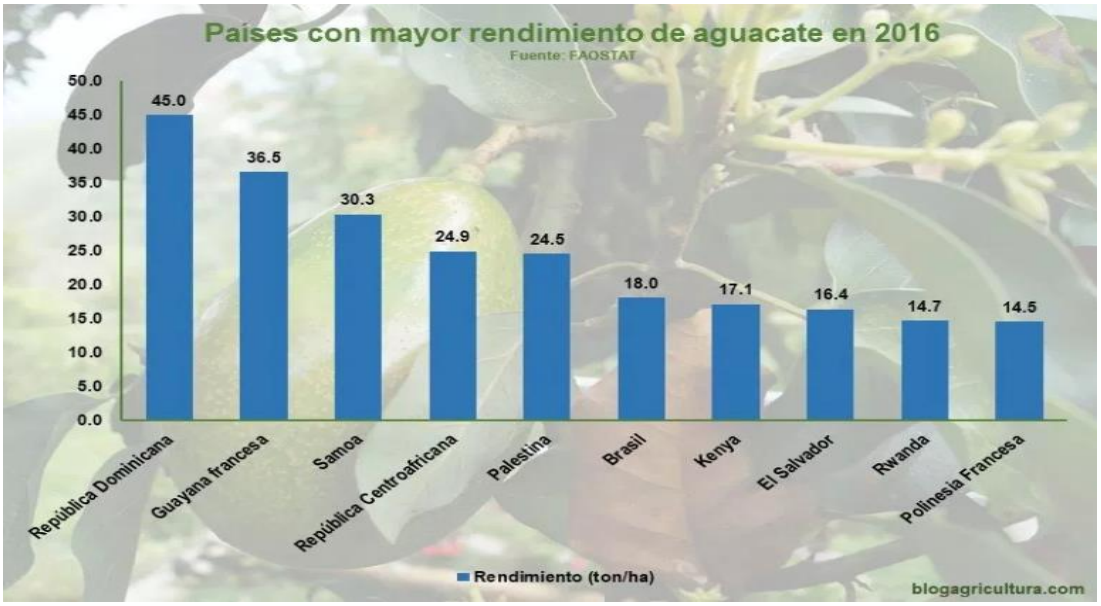
Gráfica 2. **Países con mayor cosecha de Aguacate en 2016**

Fuente: Obtenido de (FAOSTAT, 2017).

Gráfica 3. **Países con mayor producción de Aguacate en 2016**

Fuente: Obtenido de (FAOSTAT, 2017).

En términos de rendimiento, y como se puede observar en la gráfica No. 4, Colombia no hace parte de los 10 primeros países con mejor rendimiento en sus cultivos de aguacate, es decir, que a pesar que hace parte de los primeros cinco países que cosechan y producen, su rendimiento según las estadísticas presentadas y suministradas por el Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), no es el suficientemente optimo ni acorde a las condiciones que presenta Colombia para este tipo de cultivos.



Gráfica 4. **Países con mayor Rendimiento de Aguacate en 2016**

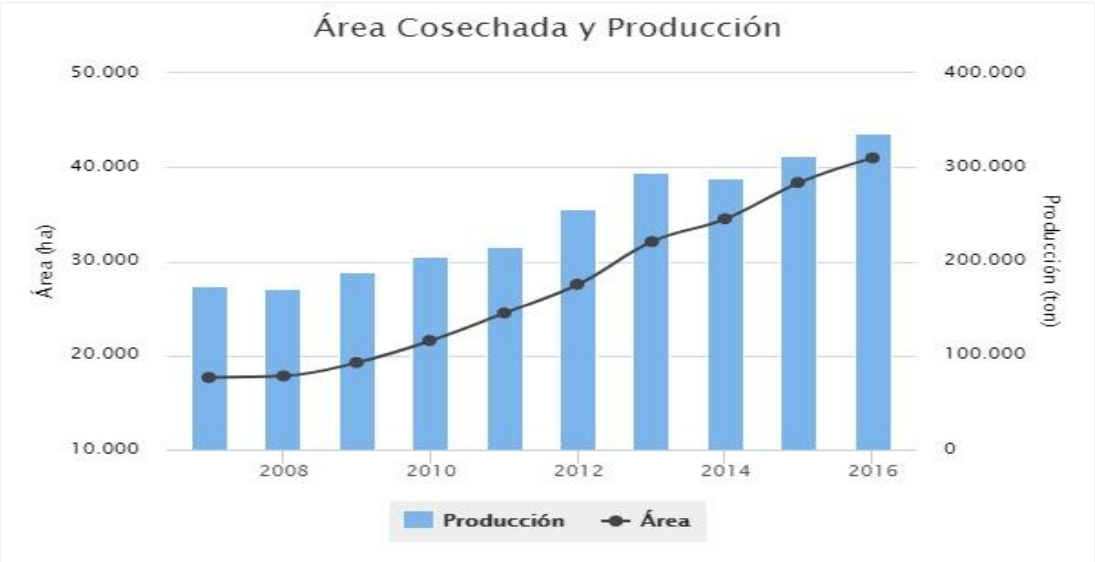
Fuente: Obtenido de (FAOSTAT, 2017).

Teniendo en cuenta la gráfica anterior y haciendo uso de las estadísticas detalladas, suministradas públicamente por la fuente Agronet, podemos observar el comportamiento que ha tenido Colombia en términos de rendimiento a través de los años, en donde se evidencia una fuerte fluctuación y una tendencia en los últimos tiempos, de que a pesar del crecimiento de la cosecha y la producción el rendimiento no crece en la misma proporción, sino por el contrario, tiende a decaer, como se observa a continuación.

Gráfica 5. **Rendimiento en cultivos de Aguacate en Colombia (2007 – 2016)**

Fuente: Obtenido de (Agronet, 2016).

Por otro lado, presentamos el comportamiento a través de los años en el periodo (2007 – 2016), en términos de cosecha y producción en cultivos de aguacate en Colombia, publicada por la misma fuente (Agronet), lo que evidencia que el rendimiento en los últimos años está siendo inversamente proporcional al área cosechada y a la producción obtenida.



Gráfica 6. **Área cosechada y producción en Colombia (2007 – 2016)**

Fuente: Obtenido de (Agronet, 2016).

Siguiendo un poco más afondo en nuestro análisis del problema y por ende en la oportunidad, acudimos a la fuente de información presentada y publicada por Finagro, en donde podemos observar el comportamiento de los cultivos de Aguacate a nivel interno en Colombia, es decir, que hasta este momento hemos analizado dichas cifras en todas las variedades de Aguacate, pero gracias a los datos suministrados por dicha fuente de información, tenemos la oportunidad de visualizar este impacto específicamente en la variedad de Aguacate Hass, como se observa a continuación.

Tabla 1. ***Datos de Siembra, Cosecha, Producción y Rendimiento en Cultivos de Aguacate de Todas las variedades Vs Hass en Colombia (2010 – 2015)***

Total nacional	2010	2011	2012	2013	2014	2015
Ha sembradas	30.006	35.373	40.510	45.070	47.762	50.292
Ha cosechadas	21.590	24.513	27.555	32.066	33.339	34.887
Toneladas	205.442	215.090	255.195	303.351	320.623	332.204
Rendimiento (ton/ha)	9,50	8,80	9,30	9,50	9,62	9,52

Estimado variedad hass	2010	2011	2012	2013	2014	2015
Ha sembradas	5.200	5.880	7.213	9.300	11.000	13.530
Participación sobre total área sembrada	17,3%	16,6%	17,8%	20,6%	23,0%	26,9%
Ha cosechadas	2.340	2.822	3.607	4.929	5.775	7.323
Participación sobre total área cosechada	10,8%	11,5%	13,1%	15,4%	17,3%	21,0%
Toneladas	23.000	26.000	29.000	35.000	47.000	58.581
Rendimiento (ton/ha)	10	9	8	7	8	8

Fuente: Obtenido de (Procolombia, 2017)

En los anteriores cuadros, se puede evidenciar que, para la variedad de Aguacate Hass, mantiene una tendencia creciente en la participación para el periodo observado (2010 – 2015), se muestra un crecimiento de siembra de 9.6 puntos porcentuales en esta variedad, un crecimiento de 10.2 puntos porcentuales en área cosechada, un crecimiento de 35,581 toneladas (Ton), y un

decrecimiento final del periodo en el rendimiento de 2 toneladas(ton) por hectárea, en la variedad de Hass.

Aterrizando un poco más la problemática al departamento del Tolima, en donde está ubicada nuestra área de estudio, vamos a observar el rendimiento a través de los años en el periodo comprendido (2007-2016) información publicada por la fuente Agronet, en donde se muestra la misma tendencia, que hemos observado a nivel nacional, pero en sus respectivas proporciones.

Gráfica 7. **Rendimiento en cultivos de Aguacate en el departamento del Tolima (2007 – 2016)**

Fuente: Obtenido de (Agronet, 2016)

El departamento del Tolima ha comprometido más área en la siembra, a su vez en la cosecha y por ende en la producción, pero su rendimiento ha sufrido un retroceso. Por lo tanto, no se está haciendo atractivo para los agricultores orientarse a este tipo de cultivos por su mayor inversión y un retorno (ROI) establecido a periodos largos, sin tener en cuenta los percances que puedan suceder durante la producción. En la siguiente tabla se puede observar en detalle la situación descrita para dicho departamento:

Tabla 2. ***Datos: Cosecha, Producción, Rendimiento y Participación en cultivos de Aguacate en el departamento del Tolima (2007 – 2016)***

Año	Area Cos. (has)	Producción (Ton)	Rendimiento (ton/has)	Participación Producción Nacional (%)	Participación Área Cos. Nacional (%)
2007	4.507	41.593	9,23	23,91	25,53
2008	4.398	41.504	9,44	24,25	24,67
2009	4.892	58.202	11,90	30,79	25,41
2010	5.835	63.475	10,88	30,90	27,03
2011	6.810	58.317	8,56	27,11	27,78
2012	5.864	51.855	8,84	20,32	21,28
2013	7.822	63.224	8,08	21,43	24,38
2014	9.054	58.649	6,48	20,31	26,23
2015	10.602	61.561	5,81	19,69	27,64
2016	10.516	58.483	5,56	17,41	25,66

Fuente: Obtenido de (Agronet, 2016).

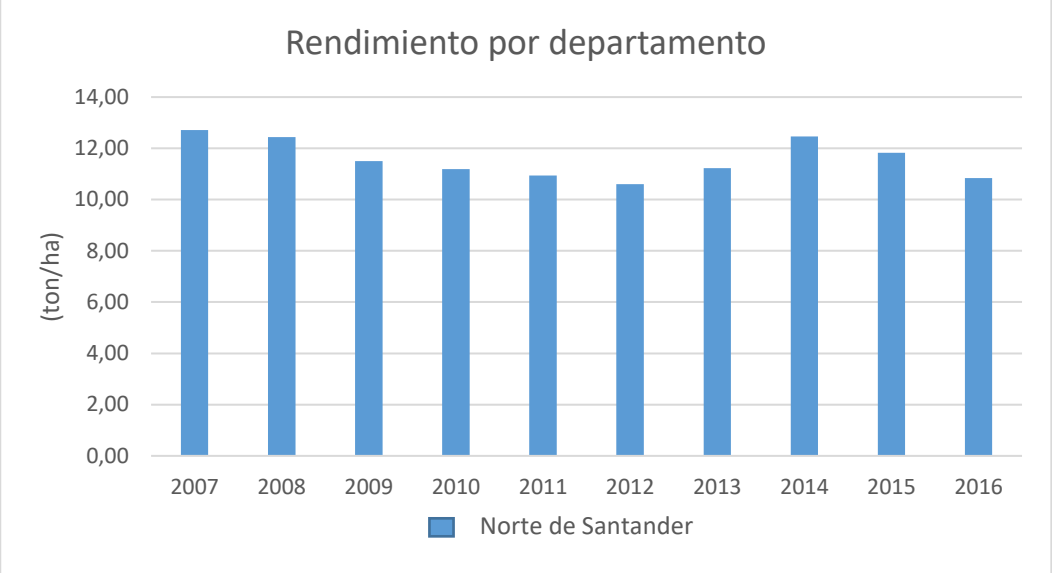
La participación de Norte de Santander en la producción nacional de aguacate aún es muy pequeña en comparación con grandes productores como Tolima o Antioquia. Sin embargo, junto con otros departamentos está adelantando proyectos de crecimiento en torno a la producción de aguacate tipo exportación (Norte de Santander, con grandes posibilidades para proveer aguacate, 2018).

Tabla 3. ***Datos: Cosecha, Producción, Rendimiento y Participación en cultivos de Aguacate en el departamento de Norte de Santander (2007 – 2016)***

Año	Area Cos. (has)	Producción (Ton)	Rendimiento (ton/has)	Participación Producción Nacional (%)	Participación Área Cos. Nacional (%)
2007	144	1.830	12,71	1,05	0,82
2008	148	1.834	12,44	1,07	0,83
2009	152	1.743	11,51	0,92	0,79
2010	146	1.634	11,19	0,80	0,68
2011	153	1.668	10,94	0,78	0,62
2012	168	1.776	10,60	0,70	0,61
2013	200	2.246	11,23	0,76	0,62
2014	350	4.362	12,46	1,51	1,01
2015	468	5.533	11,82	1,77	1,22
2016	508	5.503	10,84	1,64	1,24

Fuente: Obtenido de (Agronet, 2016).

Aunque las variaciones en el rendimiento no son abruptas y el promedio está por arriba de la media nacional, los valores son bajos en comparación con otros departamentos previendo su poca producción. Por lo tanto, se incluye como un caso de aplicación para el cual el análisis de la problemática entorno al rendimiento sirvan de referencia a otros departamentos bajo la consideración de ciertas condiciones específicas propias de la región.



Gráfica 8. **Rendimiento en cultivos de Aguacate en el departamento de Norte de Santander (2007 – 2016)**

Fuente: Obtenido de (Agronet, 2016).

1.3 Formulación de la problemática

El principal objetivo del sector agrícola es ser cada vez más competitivo con respecto al mercado nacional e internacional. Por lo tanto, debe construir estrategias para generar acciones que se consoliden en una ventaja competitiva. En este contexto, la estrategia de la producción agrícola del aguacate Hass consiste en ser más eficientes, es decir, la estrategia está enfocada en incrementar el rendimiento de los cultivos con respecto a las cifras internacionales de los países más productores y con mayor exportación. Para que un país alcance esta meta de manera sostenida

debe iniciar su estrategia desde cada unidad productiva en referencia al pequeño y gran productor agrícola. Por lo tanto, se debe encontrar la forma de aumentar el índice de rendimiento en la producción del aguacate en las organizaciones productoras agrícolas, identificando los factores que determinarán las acciones que deben implementar con decisiones apoyadas en la información y en herramientas de análisis de datos.

En este orden de ideas, es necesario ahondar en los factores que influyen en es este indicador productivo, principalmente factores los externos como las variables meteorológicas, considerando que los factores internos se encuentran controlados bajo el estudio y empleo de buenas prácticas de producción.

1.3.1 Preguntas

1.3.2 Primaria

¿Cuál es el modelo predictivo óptimo para el rendimiento de la producción de aguacate a partir de los datos meteorológicos?

1.3.3 Preguntas Secundarias

¿Qué variables meteorológicas tienen mayor influencia en el rendimiento de la producción agrícola de aguacate en Colombia?

¿Qué métodos o técnicas de análisis de datos se necesitan aplicar para la creación y evaluación del modelo?

¿Cómo integrar el modelo de análisis de datos en los procesos de la organización agrícola?

1.4 Objetivos

1.4.1 Objetivo General

Obtener un modelo predictivo que clasifique el rendimiento en la producción de aguacate a partir a los datos de las variables climáticas actuales y futuros.

1.4.2 Objetivos Específicos

- Identificar las variables meteorológicas que más influyen en el rendimiento de la producción agrícola de aguacate en Colombia.
- Determinar y aplicar los métodos y técnicas de análisis de datos para la creación del modelo predictivos.
- Desarrollar un plan de implementación de la herramienta analítica en una organización productora agrícola.
- Analizar bajo que estrategias y circunstancias el modelo propuesto puede ser integrado y puesto en producción en una organización agrícola.

1.5 Justificación

Las empresas agrícolas, más que cualquier otro tipo de empresas, necesitan hacer uso de las TIC para soportar las soluciones que se requieren al afrontar los grandes retos entorno a las estrategias y gestión de recursos que componen el proceso de tomas de decisiones. Más allá de sus distintas funciones: administrativa, financiera, técnica, comercial, de seguridad social y contable (Camón, 1963). Las empresas en el sector agrícola avistan volúmenes masivos de datos, pero desconocen cómo administrarlos de manera eficiente, útil y rápida, para lograr una mayor compresión de los factores asociados a sus actividades agrícolas y del mercado (Zikopoulos &

Eaton, 2011). Por otro lado, la actividad agrícola tradicional se ha basado en la intuición y la experiencia en tareas y calendarios específicos de siembra. Con la tecnología del Big Data como paradigma, es posible facilitar la toma de decisiones en cada momento en cuanto producción, terreno, precisión de agua y fertilizante, reduciendo los costes de producción y los niveles de contaminación para ser cada vez más sostenibles con el medio ambiente (Telefónica, 2016).

En primera instancia, es muy importante visualizar al productor agrícola como una organización con la oportunidad de desarrollar capacidades organizacionales, humanas y tecnológicas, como cualquier organización del ámbito empresarial, una vez que se acepte esta premisa podemos diseñar y adaptar las herramientas necesarias que le permitan al agricultor tomar las mejores decisiones y a su vez desarrollar cada una de dichas capacidades.

Es primordial que las decisiones de los agricultores de aguacate no se sigan tomando a través de la intuición, porque los resultados en el actual entorno dinámico no han sido acertados y han puesto en alto riesgo sus inversiones a la hora de cultivar. Lo anterior ha sido uno de los principales motivos que ha generado detrimento y temor a la hora de incentivar la producción de los cultivos en las diferentes zonas de Colombia, perdiendo la oportunidad de participar de forma competitiva en el mercado mundial del Aguacate Hass.

Desde esta perspectiva, es importante que la producción agrícola en Colombia, comience a integrar procesos de gestión soportados en los datos, para que sus decisiones sean basadas en las cifras y puedan potencializar la experiencia obtenida a través de los tiempos por los diferentes actores como entidades públicas, privadas y los mismos agricultores entre otros, que generan y siguen generando grandes volúmenes de datos, brindando la oportunidad de ser analizados para identificar patrones y tendencias de manera prospectiva, que permitan la toma de decisiones proactivas y no reactivas como se está realizando hoy en día. Adicionalmente, dichos análisis de

datos les permitirá, incrementar la confiabilidad de sus acciones y mitigar los riesgos de perdida en sus inversiones, teniendo mejores rendimientos en sus cultivos, haciéndolos costo-eficientes adoptando las mejores prácticas existentes, teniendo como base el conocimiento de las tendencias y/o patrones identificados por medio de la analítica de datos.

Para poder identificar dichos patrones fue importante la realización de un análisis retrospectivo, aprovechando los datos históricos, formulándonos una pregunta inicial, que para nuestro caso es: ¿Por qué en Colombia existe un bajo rendimiento con respecto a otros países productores de Aguacate Hass?, esta es una pregunta en la que buscamos acercarnos a la respuesta utilizando las grandes bondades de los análisis descriptivo y de diagnóstico, en los que se apoya la inteligencia de negocios para poder entender y comprender los factores que impactan el comportamiento de una variable, como lo es para nuestro caso de estudio el rendimiento.

Uno de los principales objetivos profesionales de este proyecto de investigación fue el de desarrollar el rol de Data coaching en donde su principal función es la de:

> Ocuparse de enseñar a las organizaciones 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 de negocio y a la gente de negocio a la cultura de los datos (Valencia, 2015, p.13).

Finalmente, una de las grandes razones que impulsan la realización de este proyecto fue el reto de incorporar estrategias tecnológicas al sector agropecuario, que le permita a los pequeños y medianos agricultores, el acceso a nuevas herramientas para la toma de decisiones y así puedan tener un crecimiento organizacional, económico y tecnológico estable y notable en dicho sector económico, es decir, que por medio del desarrollo de este tipo de proyectos podamos aportar

conocimiento del negocio basado en la analítica de datos apalancando el crecimiento de las capacidades del sector y por ende en el país.

1.6 Delimitación

Para el desarrollo de este proyecto se contempló la recolección de información consolidada por distintas entidades referentes en materia de producción agrícola e información meteorológica de las veinte (20) regiones con mayor participación en la producción de aguacate en Colombia. El resultado esperado consistió en un modelo estadístico que represente el comportamiento correlacionado de las variables meteorológicas con las variables estadísticas de producción.

Se realizó la recolección de información a nivel nacional, en donde se realizarán los análisis previos (Análisis de Diagnostico, Análisis Descriptivo y Análisis Predictivo).

CAPITULO 2. MARCO CONCEPTUAL

Para el desarrollo de este proyecto es importante resaltar la importancia de tener herramientas para la toma de decisiones de manera preventiva, con el fin de mitigar posibles riesgos de pérdidas financieras para el agricultor. Adicionalmente es la oportunidad de demostrar que el agro colombiano debe evolucionar utilizando metodologías y herramientas disponibles para que la toma de decisiones sea basada en los datos.

2.1 Evolución en la Toma de decisiones

Las organizaciones a través del tiempo han venido tomando sus decisiones utilizando como herramienta base la intuición es decir, la experiencia o en su defecto acuden a la herramienta clásica del juicio de expertos, grupos focales entre otros, pero teniendo en cuenta la velocidad con que va cambiando el entorno y los requerimientos del mismo, han causado que dichos métodos tradicionalistas no tengan el mismo efecto, ya que las diferentes variables del entorno en que se mueven las organizaciones no tienen un patrón definido sostenible a través del tiempo, dicho lo anterior, ha crecido la tendencia que las organizaciones tomen sus decisiones utilizando como insumo principal los datos e información que estas manejan, disminuyendo así la dependencia del instinto como herramienta principal a la hora de tomar sus decisiones, un ejemplo de ello son los sistemas de información empresarial que se encargan de capturar las grandes cantidades de datos, los cuales cada vez más las empresas recopilan y convierten en conocimiento para tomar mejores decisiones basadas en analítica de datos (Brynjolfsson, Hitt, & Kim, 2011).

Por otro lado, con la evolución de la tecnología en el mundo y especialmente en el internet ha permitido que las organizaciones y/o personas generen información con los movimientos y/o procesos que realizan en su operación diaria a través de los diferentes dispositivos, software y

sensores. El hecho que se generen diariamente grandes volúmenes de datos gestionándose de manera correcta puede generar un valor importante en el futuro de las organizaciones, marcando una tendencia de flexibilidad en su estrategia de negocio y así tenga la capacidad de responder a los nuevos requerimientos del entorno, en otras palabras, las compañías deben volcarse a entender el comportamiento de los datos para poder generar escenarios de valor y/o riqueza y así ser sostenibles en el tiempo en términos económicos, sociales y ambientales. Dentro de ese gran volumen de datos que crece estrepitosamente se gestiona una nueva era de negocios basada en datos y un cambio general de la perspectiva empresarial con decisiones enfocada cada vez más hacia los datos (Carillo, Galy, Guthrie, & Vanhems, 2018). Este nuevo tipo de organizaciones se han denominado “Organizaciones Orientadas al Dato”. Uno de los grandes retos que se enfrentan las organizaciones hoy en día es la de encontrar los mejores modelos y metodologías a la hora de analizar los datos.

2.2 Analítica de Datos

Las organizaciones a través de su historia se han realizado preguntas de manera reactiva, es decir, después de lo que ya sucedió y cuando no hay tiempo de remediar o prevenir ninguna acción o decisión, y allí es cuando los niveles de riesgo cobran relevancia, de tal manera que dependiendo el caso impactan negativamente la sostenibilidad y el crecimiento en las organizaciones. Adicionalmente con el efecto de la globalización y el dinamismo del entorno, hace que las organizaciones deban tomar decisiones de una manera más expeditas y algunas veces radicales, en donde cada vez el riesgo toma un papel más importante sobre dichas decisiones porque son tomadas desde la intuición de sus líderes.

Por la anterior razón, las organizaciones están impulsando la analítica de datos como herramienta principal para la toma de decisiones y así tener un grado de confiabilidad mitigando a toda costa el riesgo. En ese orden de ideas vamos a tomar la definición de la analítica de datos que aparece en Internet: “El Análisis de Datos (Data Analysis, o DA) es la ciencia que examina datos en bruto con el propósito de sacar conclusiones sobre la información” (Rouse, 2012), lo que quiere decir, que por medio de la analítica de datos las organizaciones adquieren conocimiento y posteriormente sabiduría de la información.

2.2.1 Niveles de Competencias Analíticas

Las organizaciones deben desarrollar diferentes tipos de capacidades y/o habilidades en los diferentes tipos de análisis para que puedan generar valor en su cadena de valor, es por eso que una organización orientada al dato debe estar en la facultad de desarrollar los siguientes tipos de análisis.

2.2.1.1 Análisis Descriptivo.

La organización es capaz de entender qué paso en las interacciones de cliente. Por ejemplo, qué han comprado o cuántos clientes han dejado de usar los servicios de la compañía (Díaz, 2016). En otras palabras, es el qué, cuál, cuánto, de una pregunta en especial, para el desarrollo de este proyecto es importante saber cuáles son los departamentos que han disminuido su rendimiento en los últimos años.

2.2.1.2 Análisis de Diagnóstico.

La organización es capaz de entender las razones por las que las interacciones con los clientes suceden. Por ejemplo, por qué los clientes compran un determinado producto (Díaz, 2016).

En otras palabras, es el por qué. Para el desarrollo de este proyecto es importante identificar por qué han disminuido su rendimiento.

2.2.1.3 Análisis Predictivo.

La organización es capaz de predecir ciertas interacciones de cliente. Por ejemplo, que clientes tienen intención de abandonar los servicios de la organización (Díaz, 2016). Adicionalmente a lo anterior, se pueden generar análisis predictivos basados en la estadística, técnicas de Machine Learning entre otros.

2.2.1.4 Análisis Prescriptivo.

La organización es capaz de tomar decisiones vinculadas con las interacciones de clientes basadas en escenarios. Por ejemplo, identificar los clientes a los que aplicar estrategias de retención (Díaz, 2016). Por ejemplo, para el desarrollo de este proyecto es importante identificar por medio de las estaciones meteorológicas del IDEAM los departamentos que no cumplan con las condiciones climáticas ideales para cumplir el rendimiento mínimo esperado.

2.2.1.5 Análisis Preventivo.

La organización es capaz de actuar con antelación a las necesidades de los clientes. Por ejemplo, enviando ofertas a los clientes ante de que se definan sus necesidades (Díaz, 2016). Para el desarrollo de este proyecto es importante tomar estrategias como, por ejemplo, tener sistemas de riegos focalizados cuando no se tenga la precipitación esperada y se tengan días más calurosos y de esa manera alcanzar los niveles de rendimiento esperado.

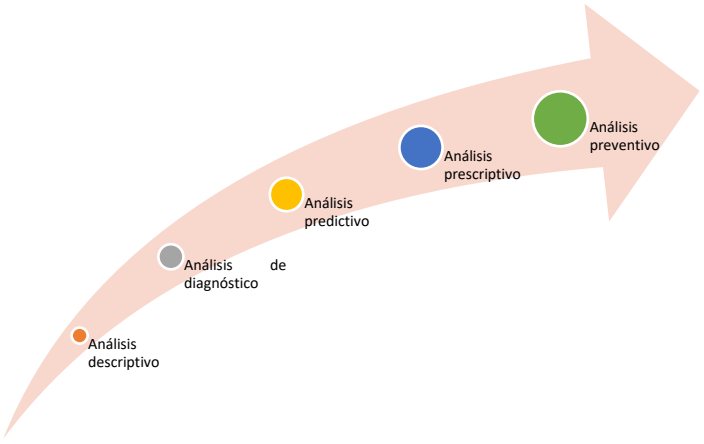


Ilustración 1. ***Niveles de competencias analíticas***

Fuente: Obtenido de (Díaz, 2016).

Los dos primeros niveles de análisis están relacionados con las herramientas de inteligencia de negocios, es decir, que su objetivo principal es gestionar los datos de una manera rápida para poder dar respuesta a las preguntas del qué, cuál, cuántos y el por qué sucedió el evento, es decir, tener a disposición los datos indicados para dar respuesta expedita aún comportamiento especial del negocio. En la siguiente ilustración se puede visualizar la interactividad entre la Inteligencia de Negocios (BI) y el análisis de datos.

Tipos de Análisis y Preguntas

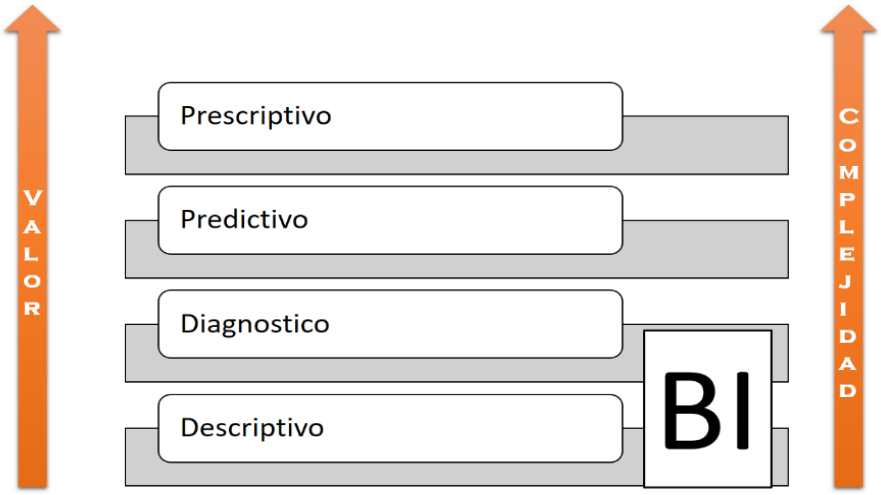


Ilustración 2. **Tipos de Análisis**

Fuente: Material académico. Obtenido de (Jorge Mario Calvo, 2018).

2.2.2 Inteligencia de Negocios (BI)

Como se observa en la ilustración No 2 la Inteligencia de negocios trabaja sobre los dos primeros análisis de la cadena (Análisis de Descriptivo y de Diagnostico), en donde está definido de la siguiente manera: Es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios. La inteligencia de negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva, que no es otra que proporcionar información privilegiada para responder a los problemas de negocio (Sinnexus, 2018)**.**

Es muy importante alinear la planeación estratégica de una organización con las preguntas a responder de la misma, para que las herramientas de Inteligencia de negocios (BI) tengan un impacto positivo en la organización y de esta manera puedan desarrollar la capacidad organizacional y tecnológico de forma estructurada, donde se pueda recorrer un camino trazado por la visión estratégica propuesta apalancando la consecución de objetivos y metas propuestas.

Ilustración 3. **Habilidades de la Inteligencia de negocios**

Fuente: Obtenido de (Sinnexus, 2018b).

2.2.2.1 Datos.

Los datos son la mínima unidad semántica, y se corresponden a elementos primarios que por sí solos son irrelevantes (Sinnexus, 2018c). Es decir, que son incipientes si no tienen ninguna aplicación de análisis, pero son la base de la información y por ende del conocimiento, por eso es importante en la analítica contemplar la identificación de fuentes de información de datos y que a su vez puedan tener un alto grado de calidad o normalización. Un ejemplo de ello, son los datos que producen las estaciones meteorológicas del IDEAM.

2.2.2.2 Información.

La información se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. Los datos se pueden transforman en información añadiéndoles valor (Sinnexus, 2018d). Una vez procesados los datos podemos generar información y determinar sí tienen o no impacto las variables meteorológicas en el rendimiento de los cultivos.

2.2.2.3 Conocimiento.

El conocimiento se deriva de la información, así como la información se deriva de los datos. Para que la información se convierta en conocimiento es necesario realizar acciones como: Comparación con otros elementos, Predicción de consecuencias, Búsqueda de conexiones, entre otros (Sinnexus, 2018e). Para el desarrollo de este proyecto es importante diseñar un modelo predictivo que permita visualizar el posible rendimiento del terreno en estudio, brindando herramientas al agricultor para tomar la decisión de cultivar o no en el terreno seleccionado. En la etapa de ejecución de este proyecto y manejando un volumen considerable de datos es importante definir los posibles *DataMart* que permitan un mejor análisis de los datos, que puedan ser almacenados en un *DataWarehouse*.

2.2.2.4 DataMart

Es un sistema orientado a la consulta, cuya distribución interna de los datos es clara y no hay dudas al respecto, estando éstos estructurados en modelos dimensionales de estrella o copo de nieve (Bigeek, 2018). Es decir, permite segmentación de la información del negocio, para el análisis de variables especificas reduciendo diferentes tipos de costos.

2.2.2.5 DataWarehouse.

Es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta (Sinnexus, 2018), es decir, una gran bodega de datos que utiliza como insumos los diferentes sistemas de información.

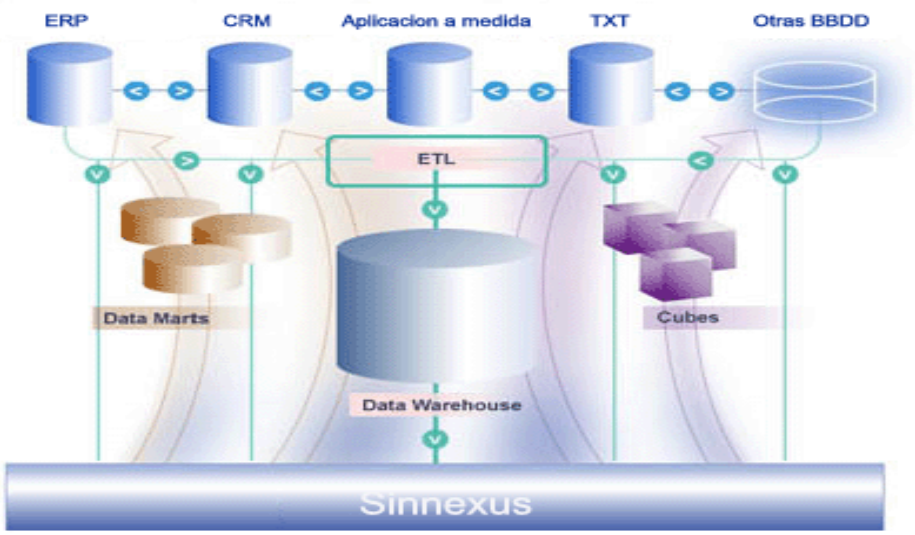


Ilustración 4. **Esquema de un DataWarehouse**

Fuente:(Sinnexus, 2018b)

2.2.3 Data Coaching

Con el gran volumen de datos que se genera hoy en el mundo, surgió un nuevo rol en las organizaciones que se denominó Data Coaching, ya que nació de la siguiente premisa: “Tener datos no nos aporta beneficios. Al contrario, nos cuestan dinero” (Valencia, 2015, p.11). En donde se ha venido percibiendo la siguiente tendencia:

Durante años, muchas organizaciones se han esforzado (y siguen haciéndolo) por capturar grandes cantidades de datos en bases de datos cada vez mayores. Además, han gastado mucho dinero en plataforma de análisis que les ofrecen muchas y potentes funcionalidades. Y, a pesar de todas las inversiones muchas organizaciones se sientes frustradas por la incapacidad de capitalizar su valor (Valencia, 2015a, p.11).

En las empresas se han venido marcando dos tipos de perfiles los que se denominan de la siguiente manera: El primer perfil son los colaboradores non-quants que es el “perfil del negocio

que debe apoyarse en la analítica" (Díaz, 2016); y el segundo perfil que son los "tecnológicos los que ayudan a realizar este tipo de proyectos (denominados quants)" (Díaz, 2016). Entre los dos perfiles descritos anteriormente existe una brecha que tiene como consecuencia un obstáculo para generar el máximo provecho a los datos generados y es por eso que surge un actor importante para reducir dicha brecha como una de las funciones del Data Coaching en donde su principal función "se ocupa de enseñar a las organizaciones 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 de negocio y a la gente de negocio a la cultura de los datos" (Valencia, 2015b, p.13). Para el desarrollo de este proyecto es importante tener en cuenta las funciones y responsabilidades del Data Coaching, ya que es importante generarle valor a los datos para el agricultor, como es el de potencializar sus cultivos materializándolo con la rentabilidad de los mismos por medio del aumento en el rendimiento de su cultivo.

Los ejes sobre los que actúa el Data Coaching son:

2.2.3.1. Cultura de Negocio.

Es importante que en la organizaciones exista un conocimiento del negocio, lo que permitirá visionarse y ajustarse a las necesidades del entorno, dicha cultura de negocio es visto por el autor de la siguiente manera: "Los datos (sistemas de información, analistas y científico de datos) no son suficientemente conscientes de lo que debe dirigir su trabajo es precisamente la búsqueda del valor económico" (Valencia, 2015c, p.18).

2.2.3.2. Cultura de Datos.

En las organizaciones es importante que las personas operacionales se involucren en el tecnicismo de la generación de los datos ya que la tendencia de las personas del "negocio no conocen el lenguaje de los datos, no están alfabetizados y no pueden aprovechar el conocimiento que les llega del lado de los datos" (Valencia, 2015d, p. 18), es importante formar perfiles profesionales integrales dentro de la organización.

2.2.3.3 Cultura Ágil.

Basados en las dos culturas anteriores, se va a permitir que "las organizaciones con un alto grado de descentralización en sus decisiones están mejor capacitadas para aprovechar las oportunidades y resolver problemas rápidamente y sobre el terreno". (Valencia, 2015, p. 18), permitiendo que sean más flexibles en su estrategia para que sean sostenibles (Económico, Social y Ambiental) a través del tiempo.

2.3 **Red Meteorológica**

Según (Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, 2005):
Una red meteorológica es el conjunto de estaciones, convenientemente distribuidas, en las que se observan, miden y/o registran los diferentes fenómenos y elementos atmosféricos que son necesarios en la determinación del estado del tiempo y el clima en una región, para su posterior aplicación a diversos usos y objetivos (p. 6).

2.3.1 Estaciones Meteorológicas

La siguiente es una clasificación de las estaciones meteorológicas basada en normas técnicas de la Organización Meteorológica Mundial, OMM y en los criterios del Instituto de Hidrología, Meteorología y Estudios Ambientales, IDEAM.

- Estación Pluviométrica (PM)
- Estación Pluviográfica (PG)
- Estación Climatológica Principal (CP)
- Estación Climatológica Ordinaria (CO)
- Estación Sinóptica Principal (SP)
- Estación Sinóptica Secundaria (SS)
- Estación Agrometeorológica (AM)
- Estación de Radiosonda (RS)
- Estación Mareográfica (MM)

De acuerdo como se clasifique una estación meteorológica contiene un tipo de instrumental como se muestra en la tabla 4.

Tabla 4. ***Tipo de Instrumental***

Tipo de Instrumental	PM	PG	CO	SS	SP	CP	AM	MM
Pluviómetro	X	X	X	X	X	X	X	
Pluviógrafo		X	X	X	X	X	X	
Sicrómetro			X	X	X	X	X	
Anemógrafo				X	X	X	X	
Heliografo					X	X	X	
Termografo					X	X	X	
Higrógrafo					X	X	X	
Tanque de Evaporación						X	X	
Actinógrafo					X	X	X	
Anemómetro						X	X	
Geotermómetros							X	
Rociógrafo							X	
Suelo (ss)							X	
Microbarógrafo				X	X			
Barómetro				X	X			
Limnímetro								X
Maxímetro								X
Limnígrafo								X
Mareógrafo								X
Salinómetro								X

Fuente: Obtenido de (Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM, 2005, p.5)

2.4 Técnicas de Predicción

2.4.1 Tipos Machine Learning

Para poder definir el modelo predictivo optimo, que nos permita generar el conocimiento objetivo de este proyecto, es importante tener en cuenta los dos tipos básicos de Machine Learning existentes:

- **Supervisado.** En este caso se tienen en cuenta en el modelo las preguntas y las características y respuestas para que se tomen como base al momento de realizar la predicción.
    - **No Supervisado.** En este caso el modelo no tiene en cuenta las preguntas ni para generar la predicción, pero si tiene en cuenta las características para que a través de las agrupaciones pueda realizar el respectivo aprendizaje.

2.5 Big Data

En la ejecución de este proyecto es posible comenzar a trabajar un gran volumen de datos suministrados por las diferentes entidades, dispositivos y sensores disponibles en los terrenos en estudio. Razón por la cual es importante visionar la ampliación de fuentes de información de fácil acceso ya que actualmente es limitado su existencia y el acceso a ella.

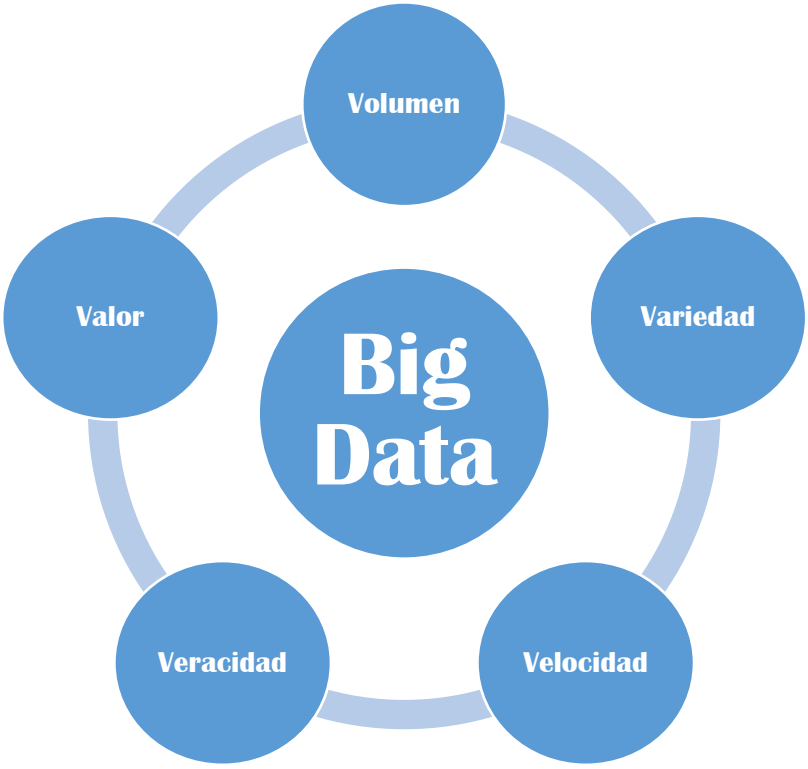


Ilustración 5. **Dimensiones del Big Data**

Fuente: Elaboración propia. Obtenida de La Teoría de IBM

Big Data surge como una nueva era en la exploración y utilización de datos. Desde la perspectiva empresarial Big Data no representa solo grandes volúmenes de datos, se deben considerar los patrones extraídos a partir de los datos y que pueden generar procesos de innovación. Desde la perspectiva tecnológica se presenta Hadoop como la principal herramienta desarrollada para el tratamiento de Big Data, incluyendo el manejo de sistemas de archivos distribuidos y el paradigma de programación Map Reduce (Leal, Méndez, & Cadavid, 2017).

2.5.1 Tipos de Datos.

Datos Estructurados. "Datos con formato o esquema fijo que poseen campos fijos. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos fundamentalmente"

(Aguilar, 2013, p. 4). Dicho tipo de datos se pueden observar en los sistemas de información tradicionales en organizaciones tales como: ERP, CRM, transaccionales entre otros.

Datos Semi-Estructurados. "Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato. Ejemplos típicos son el de texto de etiquetas de XML y HTML" (Aguilar, 2013a, p.4). Dicho tipo de datos, se pueden observar por ejemplo en los Servicios Web, el cual permite integración entre dos sistemas, en donde el HTML y el XML son lenguajes estándar que permiten la integración guardando las características e integridad de los sistemas a integrar.

Datos No Estructurados. "Ejemplos típicos de datos que no tienen campos fijos: audio, video, fotografías, o formatos de texto libre como correos electrónicos, mensajes instantáneos SMS, artículos, libros, mensajes de mensajería instantánea tipo WhatsApp, Viber, Etcétera." (Aguilar, 2013b, p.5). Dichos tipos de datos también se pueden observar en los sensores, redes sociales como Facebook, Twitter, imágenes satelitales, imágenes hyperspectral que se pueden generar con drones.

2.5.2 Big Data en el sector Agrícola y Agropecuario.

Según Serrano (2017), el Big Data y analítica brindan mejoras en el sector de la agricultura de diversas formas como por ejemplo: en inversiones a través del análisis de las tendencias; la producción de con herramientas apoyadas por la industria 4.0; transformando la logística de distribución en los procesos logísticos y optimizando la estructura de costos y en la gestión de marketing de los productos con enfoque en la segmentación de la demanda.

2.5.3 Arquitectura de Explotación de Datos

Díaz (2016), define una arquitectura en donde “cada organización, en función de sus necesidades se diseñará una arquitectura de explotación de datos personalizada. La factoría de información corporativa(FIC), está en un proceso de evolución que busca incluir las nuevas tecnologías que soportan el Big Data” (2016).

2.5.3.1 Componentes

Es importante definir cada uno de los componentes definidos en la estructura de explotación de datos, ya que es importante que las organizaciones se rijan bajo un marco de trabajo para que sea estructurado, definidos por un alcance, y por esta razón es importante definirlos de la siguiente manera:

2.5.3.1.1 Componente de Análisis

En este componente es importante destacar la “combinación de tecnologías de *Business Intelligence,* analítica e inteligencia operacional que busca generar valor a partir del dato” (Díaz, 2016). Teniendo en cuenta que la sabiduría parte del conocimiento, el conocimiento parte de la información y la información parte del dato.

2.5.3.1.2 Componente de gestión del dato

En este componente es importante definir el gobierno de los datos y así “Busca mejorar la gestión del dato en la organización desde todas sus facetas: seguridad, accesibilidad, disponibilidad, consistencia, etc.” (Díaz, 2016).Para esto, es importante definir el alcance y cada uno de los roles de los actores involucrados.

2.5.3.1.3 Componente de Procesamiento y almacenamiento

En este componente es importante definir el framework seleccionado para el procesamiento de datos teniendo en cuenta la "Combinación de la FIC con tecnologías de Big data, como Hadoop y otras tecnologías NoSQL, in-memory o MPP" (Díaz, 2016).

2.5.3.1.4 Componente de fuentes

En este componente se "considera todas las fuentes de información a explotar tanto interna (como sistemas transaccionales, sistemas de información, corporativa, etc.) como externa (redes sociales, sensores, etc.)" (Díaz, 2016). Es decir, los diferentes tipos de dato especificados anteriormente.

A continuación, se puede observar la arquitectura propuesta por el autor

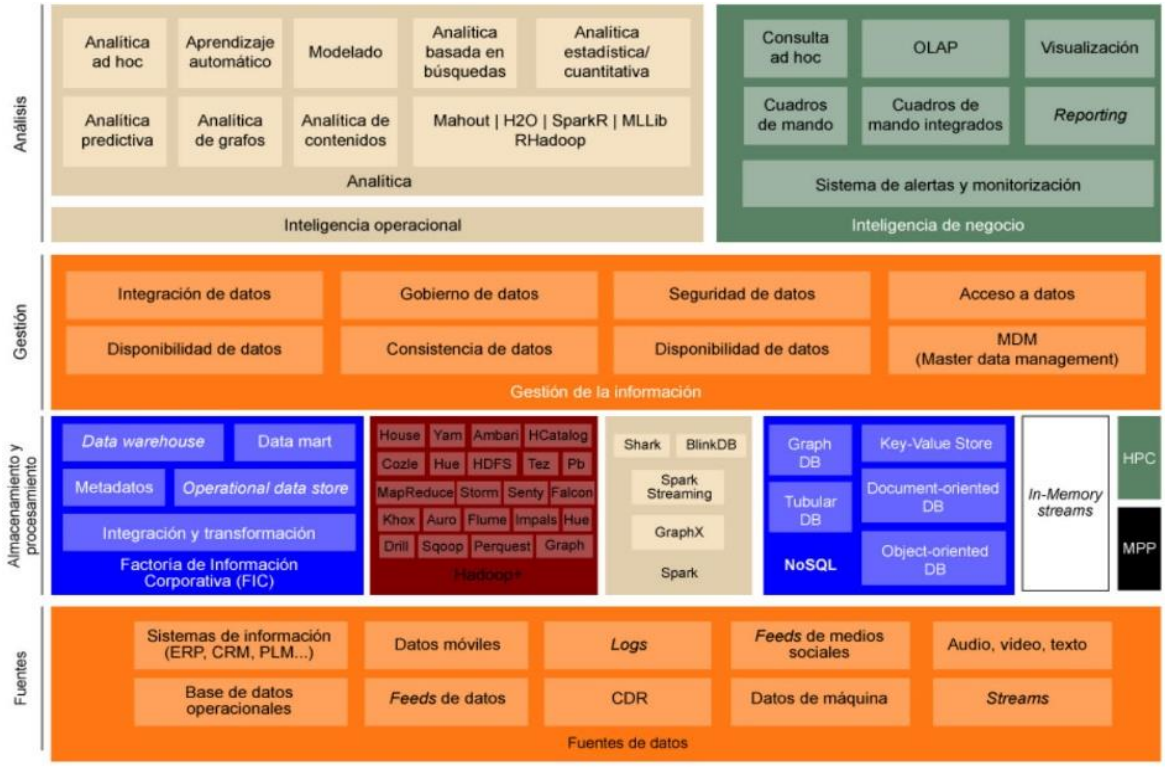


Ilustración 6. **Plataforma de generación de valor a partir del dato**

Fuente: Obtenido de (Díaz, 2016).

2.6 Fuentes de Información

Existen diferentes tipos de fuentes de información que provienen de diferentes formas: Entidades publica, privadas, redes sociales, audios, videos, imágenes, satélites; que permiten generar datos incipientes, que aplicándoles análisis podemos generar información y posteriormente conocimiento.

2.6.1 GPS

El Sistema de Posicionamiento Global (GPS) "es un sistema de radionavegación de los Estados Unidos de América, basado en el espacio, que proporciona servicios fiables de posicionamiento, navegación, y cronometría gratuita e ininterrumpidamente a usuarios civiles en todo el mundo" (GPS, 2018). El GPS se emplea como herramienta para brindar información exacta sobre el posicionamiento en el terreno y levantamientos topográficos que orienten sobre los relieves de la tierra y otros datos que contribuyen en la mejora de la planificación en el marco de la agricultura de precisión (Agricultura Moderna, 2018).

2.6.2 Drones

Los drones son una de las herramientas más novedosas aplicadas a la agricultura para realizar tareas de fumigación, siembra, polinización y monitoreo de plagas y suelos entre otras aplicaciones. Por definición un Drone es todo aquel vehículo aéreo manejado de forma remota: "La mayor evolución en cuanto a las técnicas en agricultura de precisión viene del desarrollo de nuevos sensores y su aplicación extensiva mediante los drones" (AINIA, 2015).

2.6.3 Internet de las Cosas (IoT)

Según Posada (2017), “El Internet de las Cosas se refiere a un sistema interrelacionado de dispositivos de cómputo, máquinas digitales o mecánicas, objetos, animales o personas los cuales poseen un identificador único y la habilidad de transferir datos sobre una red sin requerir interacción humana” (p. 15).

La mayoría de las empresas monitorean los flujos de información separados. Con IoT, es posible aplicar la trazabilidad a los productos a través de identificadores únicos, vinculando la información generada con su procedencia, uso y destino (Strange & Zucchella, 2017).

La agricultura requiere la captura, almacenamiento y análisis de los datos proveniente de la red sensores, en el contexto del internet de las cosas estarán interconectado en una misma red para obtener acceso a la información de forma remota.

2.6.4 Sensores

Sensor o transductor es un dispositivo capaz de obtener una señal física y transformarla en eléctrica para que sea capturada, almacenada y analizada. Los sensores en tecnologías físicas y químicas para medir variables como la humedad, el pH del suelo, el nivel del agua, temperatura y muchas más de acuerdo a la aplicación (máquinas agrícolas, en el campo y el aire) y tecnología utilizada puede variar su forma, tamaño y ubicación, cumpliendo con el objetivo proyecto a la eficiencia ya sea reduciendo costo, mejorando la productividad o reduciendo el impacto medio-ambiental (Bogue, 2017).

2.6.5 Datos Abiertos

Datos abiertos: son datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las

entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos (DNP, 2018, p. 97).

2.7 Gestión de riesgos

Los beneficios que proporciona la gestión de datos dentro de las aplicaciones no son indiferentes al sector al cual aplica. Por lo tanto, los beneficios que percibe las entidades financieras en gestión del riesgo a través de la analítica de la información son también aplicables al sector de la agricultura en el marco de la herramienta en la gestión de riesgos para los productores.

2.7.1 Riesgos positivos (oportunidades)

Los riesgos se asocian a cada tipo de empresa y al sector en el que desarrolla su actividad económica. Con la gestión de la información es posible prescribir riesgos estratégicos que conduzcan a una oportunidad o ventaja competitiva.

2.7.2 Riesgos negativos (amenazas)

La gestión de datos con aplicaciones para el uso del Big Data constituye una poderosa herramienta para mitigar los riesgos presentes en los procesos productivos de la agricultura y soportan la estrategia para prescribir y prevenir situaciones adversas que amenazan la continuidad del negocio.

2.8 Conceptos y tipos técnicas estadísticas

- **Análisis de la Varianza (ANOVA):** Es una herramienta estadística para detectar diferencias entre significativos grupos experimentales (Sawyer, 2013).

- **Chi-Cuadrado:** La estadística de chi-cuadrado es una herramienta no paramétrica (sin distribución) diseñada para analizar las diferencias de grupo cuando la variable dependiente se mide a un nivel nominal. Como todas las estadísticas no paramétricas, el Chi cuadrado es robusto con respecto a la distribución de los datos (McHugh, 2013).
- **Correlación:** La correlación es un método estadístico utilizado para evaluar una posible asociación lineal entre dos variables continuas. Es simple tanto para calcular como para interpretar (Mukaka, 2012).
- **DBSCAN** (Clustering espacial basado en densidad de aplicaciones con ruido): Es un método basado en densidad, que encuentra objetos centrales, es decir, objetos que tienen vecindarios densos. Conecta los objetos centrales y sus vecindarios para formar regiones densas como clústeres (Jiawei Han, Micheline Kamber, 2011).
- **Medoides:** Es una medida de tendencia central. Es el punto desde donde la suma total de la distancia al cuadrado de todos los puntos es el mínimo (Maheshwari, 2015).
- **PAM (A**lgoritmo de partición alrededor de los medoides): Es un algoritmo basado en particiones. Divide los datos de entrada en número de grupos. Encuentra un conjunto de objetos llamados medoides que están ubicados centralmente. Con los medoides, los puntos de datos más cercanos se pueden calcular y convertir en clústeres (Jiawei Han, Micheline Kamber, 2011).
- **CLARA** (Aplicaciones de grandes clústeres)**:** Para tratar con conjuntos de datos más grandes, se puede usar un método basado en muestreo llamado CLARA (Jiawei Han, Micheline Kamber, 2011). En lugar de encontrar objetos representativos para todo el conjunto de datos, CLARA extrae una muestra del conjunto de datos, aplica PAM en la muestra y encuentra los medoides de la muestra (Jiawei Han, Micheline Kamber, 2011).

- **Redes neuronales artificiales:** cuando se utiliza para la clasificación, suele ser una colección de unidades de procesamiento de tipo neuronal con conexiones ponderadas entre las unidades (Jiawei Han, Micheline Kamber, 2011).
- **Redes neuronales no paramétricas:** las redes neuronales no paramétricas toman la posición de que el número y la naturaleza de los parámetros son flexibles y no se fijan de antemano (Jiawei Han, Micheline Kamber, 2011). A diferencia de las redes neuronales paramétricas, la dimensionalidad de las matrices de peso es indeterminada (Philipp & Carbonell, 2017).

2.8.1 Tipos de correlación

- **Correlación del momento del producto de Pearson:** Se considera una correlación lineal simple, lo que significa que la relación entre dos variables depende de que sean constantes. Pearson se usa con datos de intervalo para medir la fuerza de una correlación (Kerley, 2017).
- **Correlación de rango de Spearman:** La ecuación de Spearman es más simple y se usa a menudo en las estadísticas en lugar de Pearson, aunque es menos concluyente. La correlación se calcula utilizando una hipótesis nula que se acepta o rechaza posteriormente (Kerley, 2017a).
- **Correlación de rango de Kendall:** mide la fuerza de dependencia entre los conjuntos de dos variables aleatorias. Kendall puede usarse para un análisis estadístico adicional cuando la Correlación de Spearman rechaza la hipótesis nula. Alcanza una correlación cuando el valor de una variable disminuye y el valor de la otra variable aumenta; esta correlación se conoce como pares discordantes (Kerley, 2017b).

- **Hipótesis nula:** En las estadísticas inferenciales, la hipótesis nula es una afirmación general o una posición predeterminada que indica que no existe una relación entre dos fenómenos medidos o una asociación entre grupos (Agrawal, 2019).
- **Hipótesis alternativa**: Es la hipótesis utilizada en la prueba de hipótesis que es contraria a la hipótesis nula. Se suele considerar que las observaciones son el resultado de un efecto real (con una cierta cantidad de variación de azar superpuesta) (Agrawal, 2019a).
- **Bosque aleatorio** (Random Forest): El bosque aleatorio (RF) es una técnica de "aprendizaje conjunto" que consiste en la agregación de un gran número de árboles de decisión, lo que resulta en una reducción de la varianza en comparación con los árboles de decisión individuales (Couronné, Probst, & Boulesteix, 2018).
- **Validación de espera** (Hold-out validation): Consiste en separar aleatoriamente y mantener el conjuntos de datos en una parte más grande para el entrenamiento y una más pequeña validar los resultados (Zheng, 2015).
- **Validación cruzada** (cross validation): la validación cruzada es una técnica común para estimar la calidad (precisión) de un clasificador inducido por algoritmos de aprendizaje supervisado (Ullrich, 2009).
- **Validación cruzada K-Fold** (*K-fold cross validation*): Es un tipo de validación cruzada que se utiliza cuando no tenemos suficientes datos para reservar un conjunto de validación. Consiste en dividir el conjunto de datos en K particiones del mismo tamaño y el modelo se entrena para cada partición creada con el restante de particiones hasta cubrir todas las particiones y obtener el resultado final que es la media de todos los resultados (Hastie, Tibshirani, & Friedman, 2009).

- **Regresión lineal:** Es una técnica estadística para investigar y modelar la relación entre dos variables (Montgomery, Peck, & Vining, 2015)
- **Regresión logística:** La regresión logística mide la relación entre una variable dependiente categórica y una o más variables independientes (Maheshwari, 2015)
- **Múltiple regresión lineal:** La regresión lineal múltiple es una variación de la técnica estadística de regresión lineal. Es usado cuando se quiere determinar el modelo lineal más apropiado para predecir solo una variable de fondo dependiente y a partir de un conjunto de variables independientes observadas (Timm, 2007).
- **Matriz de confusión:** Una matriz de confusión o matriz de error, es una tabla de contingencia que sirve como herramienta estadística para el análisis de observaciones emparejadas. El contenido de una matriz de confusión es un conjunto de valores que contabilizan el grado de semejanza entre observaciones emparejadas (Ariza López, Rodríguez Avi, & Alba Fernandez, 2018).

Se utiliza para calcular la precisión de la predicción clasificando los puntos de datos en una matriz entre una clase actual versus la predicción. Denominando verdaderos positivos y verdaderos negativos a los valores predichos que coinciden con los actuales. Por otro lado, se denomina falso positivo y verdadero negativo a los que no coinciden. El valor más alto puede ser del 100 por ciento. En la práctica, los modelos predictivos con más del 70 por ciento de precisión pueden considerarse usables en dominios de negocios, dependiendo de la naturaleza del negocio (Maheshwari, 2015)

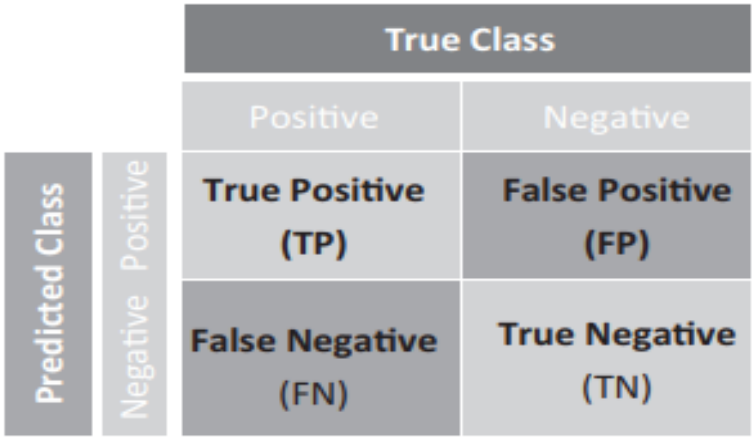


Ilustración 7. **Matriz de Confusión**

Log Loss=Log-loss es una medida de precisión que incorpora la idea de confianza probabilística. Tiene en cuenta la incertidumbre de su predicción en función de cuánto varía de la etiqueta real (Swalin, 2018).

CAPITULO 3. MARCO METODOLÓGICO

3.1 Metodología

En el marco de gestión de los objetivos del sector y un enfoque de los productores como organización agrícola, se analiza una metodología de minería de datos que tenga como primer objetivo la comprensión del negocio y permita integrar la tecnología con la estrategia organizacional a través del conocimiento generado por un eficiente uso de los datos. las metodologías más usadas en los proyectos de análisis y minería de datos, como lo demuestra los resultados de una encuesta publicada por el sitio web Kdnuggets(Piatetsky, 2014), en donde realiza

una interesante comparación de las diferentes guías de referencia entre el año 2007 y 2014 como se puede observar en la Ilustración No. 8.

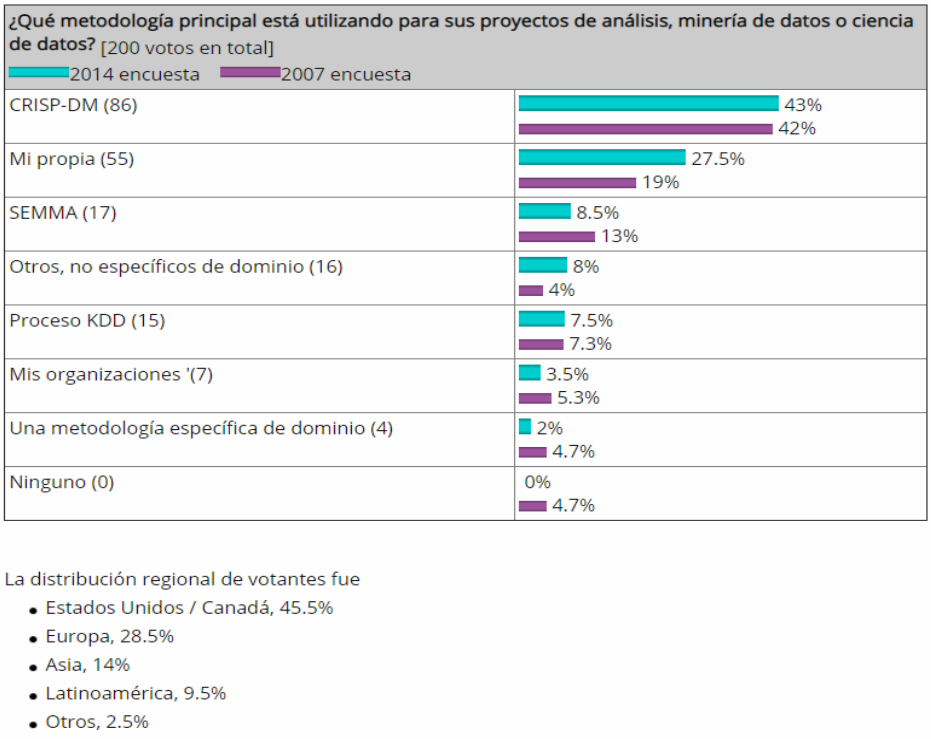


Ilustración 8. **Resultados de encuesta de metodologías para la minería de datos más utilizadas**

Fuente: Obtenido de (Piatetsky, 2014).

De acuerdo a las características previamente presentadas se opta por hacer uso de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). En relación con la respuesta de los autores en su artículo "Metodología Crisp para la implementación Data Warehouse" publicado en el año 2009, a la pregunta del por qué utilizarla "La Metodología CRISP está sustentada en estándares internacionales que reflejan la robustez de sus procesos y que facilitan la unificación de sus fases en una estructura confiable y amigable para el usuario. Además de ello esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y reciproco." (Salcedo Parra, Galeano, & Rodriguez, 2010), lo anterior, les garantiza a las organizaciones un dinamismo evolutivo en la generación de

conocimiento, basado en los objetivos de negocio. La metodología CRISP-DM se compone de 6 fases como se puede observar en la Ilustración 9.

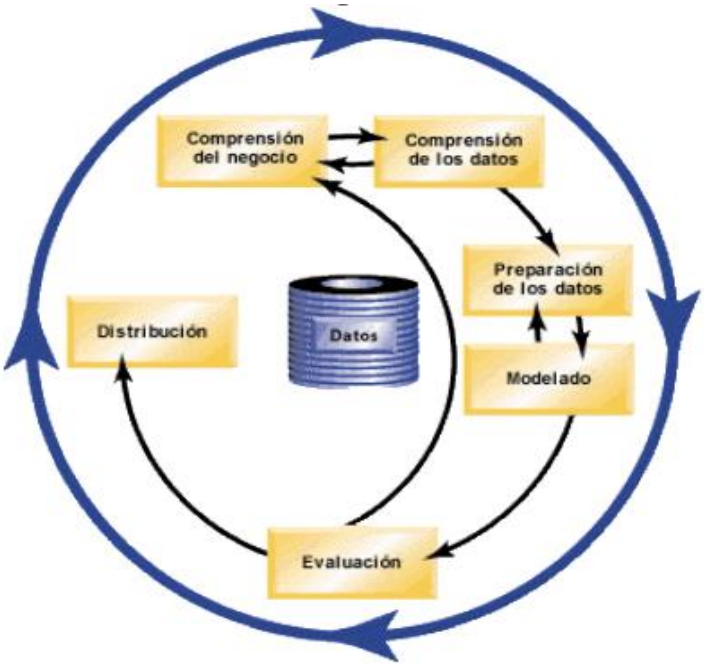


Ilustración 9. **Fases de la Metodología CRISP-DM**

Fuente: Obtenido de (IBM, 2011).

3.1.1 Fase 1: Comprensión del negocio

Esta fase es una de las más importantes de la metodología en donde se debe tener la capacidad o habilidad de formular la pregunta correcta para aplicar el proceso de analítica de datos. Para ser exitoso en la formulación de dicha pregunta se debe conocer y entender la organización desde un punto de vista estratégico, es decir, siendo congruentes con el planteamiento estratégico de la organización y la visión del proyecto a desarrollar.

Adicionalmente es necesario definir los objetivos que motivara la realización del proyecto de analítica de datos propuesto en la organización. Para el desarrollo de esta fase es importante tener en cuenta el punto de vista que tuvo el ingeniero: José Alberto Gallardo Arancibia, en el

desarrollo de su proyecto de grado, denominado:" Metodología para la Definición de Requisitos en Proyectos de Minería de Datos (ER-DM)", donde hace mención a lo siguiente: Para obtener el mejor provecho de Minería de Datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Minería de Datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio (Gallardo, 2009, p. 17).

A continuación, podemos observar las actividades de la metodología en la fase de compresión del Negocio.

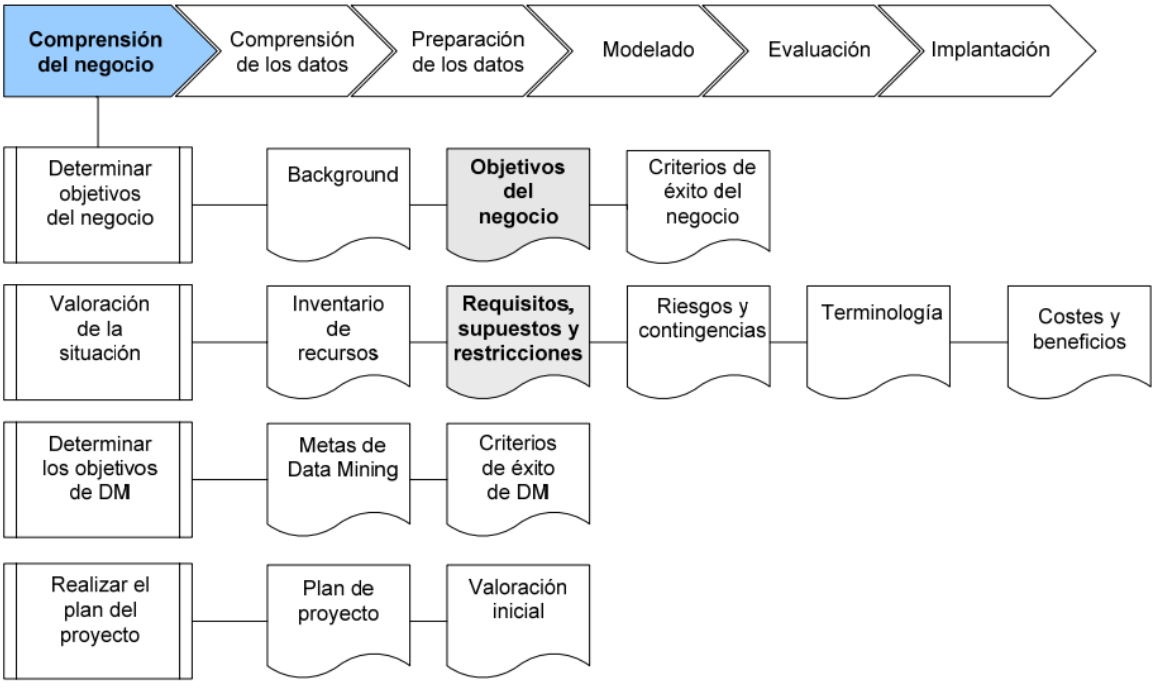


Ilustración 10. **Fase Comprensión del Negocio – Actividades**

Fuente: Obtenido de (José & Gallardo Arancibia, 2009, p. 17).

Como se observa en la ilustración anterior existen 4 actividades en donde se definen de la siguiente manera:

- *Determinar objetivos del negocio:* En esta actividad se definen los objetivos de la organización tipo SMART (Especifico, medible, alcanzable, orientado a resultados, fecha límite de ejecución), en donde se puedan determinar los criterios de éxito de la organización.
- *Valoración de la Situación:* En esta actividad se realiza un análisis del entorno, en donde se tiene en cuenta (Recursos, riesgos, restricciones, costos, etc.). Adicionalmente se plantean supuestos o hipótesis iniciales y se determinan los costos y beneficios.
- *Determinar los objetivos de la Analítica de Datos:* En esta actividad, se determinan los objetivos tipo SMART (Especifico, medible, alcanzable, orientado a resultados, fecha límite de ejecución) del proceso de Analítica de datos a realizar, en donde, se definen aquellos criterios de éxito que van a determinar la calidad del proyecto propuesto.
- *Realizar el plan del Proyecto*: En esta actividad se obtiene un entregable bastante importante en donde se puede visualizar el plan de actividades del proyecto a realizar.

Para la aplicación del presente proyecto de grado en los dos primeros capítulos se pueden observar los objetivos, el análisis del entorno actual, y en la ejecución se observará el plan de actividades del proyecto a realizar con el análisis de datos, y de esta manera cumplir con las actividades planteadas por esta fase. Por otro lado, se plantea una hipótesis inicial basada en las diferentes lecturas realizadas y en los datos presentados por varias entidades en donde se consideramos lo siguiente: “El principal factor que ha impactado negativamente los cultivos de Aguacate Hass en el territorio nacional, es el Cambio Climático, principalmente por los fenómenos

que tuvieron presencia desde el año 2010", basados en esta premisa planteamos el desarrollo del presente proyecto de Analítica de datos, considerando las variables relevantes impactadas por dicho tipo de fenómenos.

3.1.2 Fase 2: Comprensión de los datos

En esta fase se comienzan a reunir los recursos para llevar a cabo el proceso de analítica de datos propuesto, es decir, basado en uno de los entregables de la fase anterior, como lo es la hipótesis inicial, se identifican aquellas posibles variables que tienen impacto en la pregunta formulada en dicha fase, y una vez postuladas estas variables, se identifican los medios de recolección de la información necesaria en los periodos requeridos por cada una de ellas. Lo anterior, es el insumo principal para evaluar la correlación entre cada una de las variables propuestas y de esta manera poder validar la hipótesis inicial planteada. A continuación, podemos observar las actividades propuestas por la metodología en la fase de comprensión de los datos.

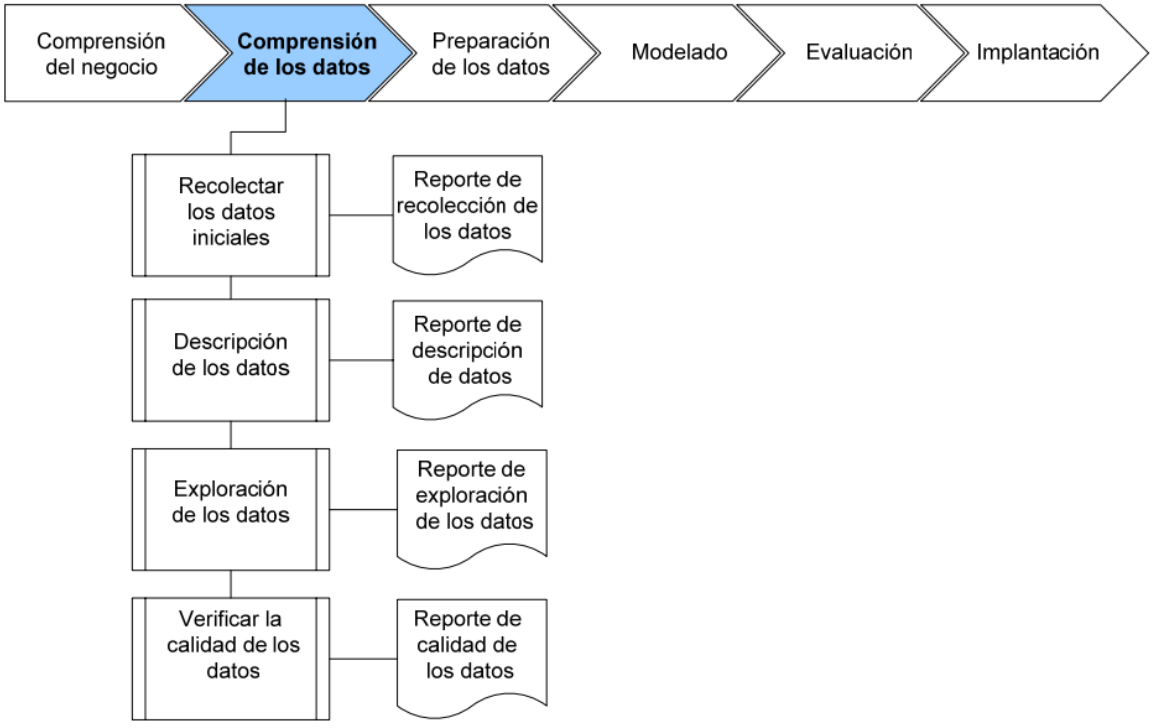


Ilustración 11. **Fase Comprensión de los Datos – Actividades**

Fuente: Obtenido de (José & Gallardo Arancibia, 2009)

- **Recolectar los datos iniciales:** En esta actividad se genera un listado de los datos que se deben recolectar por cada una de las variables postuladas, en donde, se define lo siguiente: La fuente que suministra la información, la periodicidad de la información.
- **Descripción de los datos:** una vez ya postuladas las variables se debe construir la descripción de los datos a recolectar, como lo es: Unidad de medida, formato de reporte del dato por la fuente suministrada, esta actividad se debe representar por medio de un listado.
- **Reporte de exploración de los datos:** En esta actividad a los datos recolectados se debe aplicar estadística básica para poder tener un panorama de la información adquirida como, por ejemplo: (Máximos y Mínimos, Media, etc.), que nos permitan tener un panorama claro de los datos recolectados, esta actividad se debe representar por medio de un componente gráfico.
- **Reporte de calidad de los datos:** Una vez que se tienen los datos se debe revisar la calidad de los mismos y postular los procesos de normalización que se deberían aplicarse para cada uno de ellos con el objetivo de contar con datos de buena calidad para que el proceso modelado tenga un buen nivel de confiabilidad, esta actividad se debe representar por medio de un componente gráfico o en su defecto un listado.

Para el presente proyecto de grado y tomando como base la hipótesis inicial propuesta, se postulan las siguientes variables que impactan los cultivos de Aguacate Hass, debido al cambio climático, para el posterior proceso de recolección de datos:

- *La precipitación*: Es agua líquida o sólida que cae de las nubes a la superficie de la Tierra o se forma en diferentes cuerpos como resultado de la condensación del vapor de agua en la atmósfera. La precipitación puede ser líquida, sólida o mixta. La precipitación líquida incluye lluvia y llovizna (Borzenkova, 2009), su unidad de medida a utilizar es: milímetro (mm).
- *Temperatura:* Es el grado de calentamiento de un cuerpo (o su estado térmico) medido con la ayuda de un instrumento especial, un termómetro (Melker, Starovoitov, & Vorobyeva, 2010)**.** La unidad de medida a utilizar es: grados Celsius (ºC).
- *Humedad Relativa:* Se trata de la forma más corriente de expresar la humedad atmosférica. Es la relación entre el contenido de vapor en un momento determinado y el máximo que podría contener si estuviese saturado (Fariña, 1998). Se expresa en tanto por ciento (%)
- *Tensión de vapor:* Es la presión parcial ejercida por el vapor de agua contenido en el aire. Cada uno de los gases que forman la atmósfera es responsable de una parte de la presión atmosférica (o tensión atmosférica), y el vapor de agua también (Fariña, 1998). Se expresa en milibares(mb).
- *Punto de rocío:* Indica la temperatura a la que se saturaría la cantidad de vapor existente actualmente en el aire… su variación diaria es mucho menor que cualquiera de los demás parámetros. Como se puede comprender, punto de rocío y tensión de vapor están muy relacionados (Fariña, 1998). Se expresa en grados Celsius (ºC).

La fuente de información gestionada para la adquisición de los datos de cada una de las variables anteriores es el “Instituto de Hidrología, Meteorología y Estudios Ambientales” (IDEAM), el cual cuenta con una red meteorológica en todo el perímetro nacional que tiene la capacidad de realizar estas mediciones en diferentes periodicidades.

3.1.3 Fase 3: Preparación de los datos

Una vez recolectados los datos en la periodicidad necesitada, se procede a preparar los datos para el modelo, en donde se aplican los procesos de normalización propuestos para la fase anterior para lograr el nivel de calidad deseado, teniendo como premisa lo siguiente: “a mayor calidad de los datos recolectados, mayor es la confiabilidad del modelo de análisis predictivo”. Adicionalmente se debe recordar que la fase siguiente es la fase de modelado, en donde se debe tener en cuenta lo siguiente: Dependiendo del modelo de predicción a utilizar deben preparar los datos, ya que cada modelo tiene unos requerimientos de alistamiento de los datos diferentes. A continuación, podemos observar las actividades propuestas por la metodología en la fase de preparación de los datos.

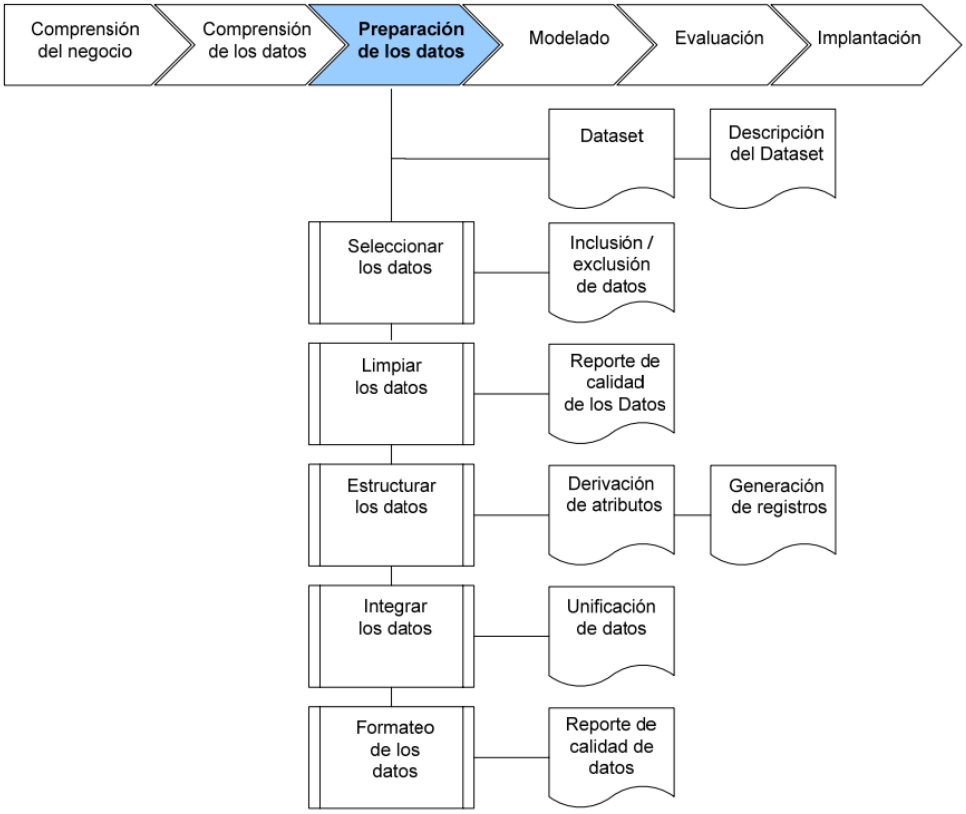


Ilustración 12. **Fase Preparación de los Datos – Actividades**

Fuente: Obtenido de (José & Gallardo Arancibia, 2009)

- *Seleccionar los datos:* En esta actividad se debe realizar un proceso de inclusión y/o exclusión de los datos, con el objetivo de utilizar los datos que cumplan los criterios de aceptación del modelo a utilizar, con el fin de descartar aquellos datos que puedan generar "ruido" al modelo y por ende afecte la confiabilidad del mismo. El entregable de esta actividad es un **Data set**.
- *Limpiar los datos:* En esta actividad se aplican los respectivos procesos de normalización para obtener el mayor nivel de calidad esperada y de esta manera buscar un buen grado de confiabilidad del modelo predictivo, el Entregable de esta tarea serán componentes gráficos de estadística básica aplicada.

- *Estructurar datos:* En esta actividad se aplican los procesos de conversión de datos a diferentes escalas, o generación de nuevos atributos que representen la transformación de los datos por las condiciones propuestas, según los requerimientos del modelo predictivo.
- *Integrar los datos:* En esta actividad se pueden unificar datos de diferentes tablas en nuevos atributos, según las necesidades del modelo predictivo.
- *Formateo de los datos:* En esta actividad se pueden aplicar diferentes procesos como ordenación, agrupación que faciliten la aplicación del modelo predictivo a utilizar.

Para el desarrollo del presente proyecto, y específicamente en la aplicación esta fase, el proceso de normalización va a ser focalizado ya que los datos son estructurados, posiblemente se tengan que aplicar procesos de formateo de datos y/o de estructura dependiendo a la aplicación del modelo predictivo.

3.1.4 Fase 4: Modelado

Esta fase es una de las más importantes, ya que en este momento es donde se va a seleccionar y a ejecutar la técnica del modelo predictivo a utilizar. A continuación, podemos observar las actividades propuestas por la metodología en la fase de modelado.

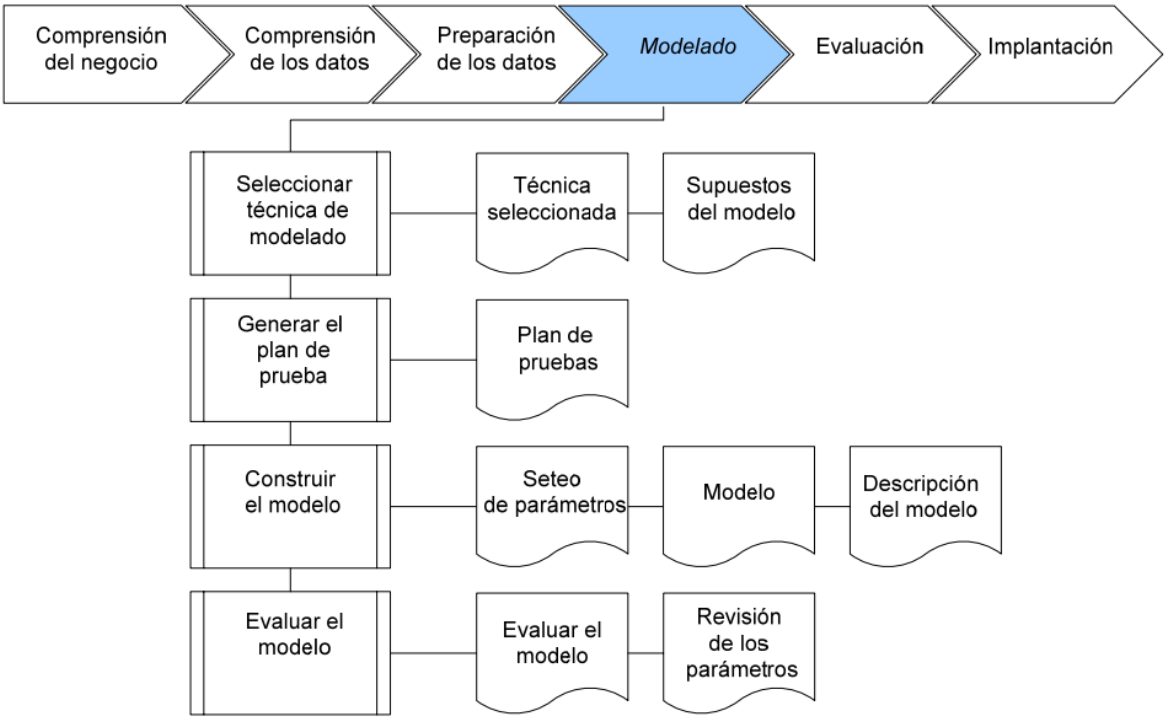


Ilustración 13. **Fase Modelado – Actividades**

Fuente: Obtenido de (José & Gallardo Arancibia, 2009)

- *Seleccionar técnica de modelado:* en esta actividad se debe seleccionar la técnica del modelo y para su selección compartimos los criterios del ingeniero José Alberto Gallardo Arancibia, que menciona durante del desarrollo de su proyecto de grado.
    - Ser apropiada al problema
    - Disponer de datos adecuados
    - Cumplir los requisitos del problema
    - Tiempo adecuado para obtener un modelo
    - Conocimiento de la técnica
    - Basado en lo anterior se debe seleccionar el modelo de acuerdo con el objetivo principal del proyecto, como se observa en la siguiente tabla (Gallardo, 2009, p. 21).

Tabla 5. ***Técnicas de modelado según clasificación***

Objetivo de clasificación	Tipos de técnicas de modelado
Clasificación	Árboles de decisión (DT) K-nearest neighbour (KNN) Razonamiento basado en casos (CBR)
Predicción	Análisis de regresión (RA) Redes Neuronales (NN)
Segmentación	Redes Neuronales (NN) Técnicas de visualización (nube de etiquetas, visualizaciones interactivas, etc.)

Obtenido de (José & Gallardo Arancibia, 2009, p. 21)

.

- *Generar el plan de pruebas:* En esta actividad una vez establecido o seleccionado el modelo se debe definir un plan de pruebas en donde se puedan comprobar los resultados una vez el modelo de predicción propuesto se haya ejecutado. Adicionalmente se debe contemplar las posibles actividades de aplicación de los respectivos ajustes en el caso que se necesiten, el entregable de esta actividad es el plan de actividades para las pruebas de comprobación de resultados.
- *Construir el modelo:* En esta actividad se lleva a cabo la construcción del modelo basado en la selección correspondiente, como se explica en actividades anteriores y según la preparación de los datos recolectados, el entregable de esta actividad ya es el algoritmo o el diagrama del modelo desarrollado.
- *Evaluación del modelo:* Para evaluar el modelo desarrollado es importante que se tenga el conocimiento del negocio para evaluar los resultados obtenidos. En el caso particular del proyecto, los parámetros de producción agrícola. Generalmente las interpretaciones están dadas, por un lado, al criterio preexistente de los ingenieros en minería de datos y por el otro, en los expertos en el contexto del dominio del tema (Gallardo, 2009, p. 22)

3.1.5 Fase 5: Evaluación

En esta fase de debe evaluar los resultados del modelo ejecutado en la fase anterior, y una de las maneras más pragmáticas de hacerlo es tomando un grupo de registros históricos que no se tuvieron en cuenta en el desarrollo del modelo, para realizar las diferentes pruebas de evaluación, lo anterior con el fin de poder comprobar el modelo desarrollado Vs la realidad y de esta manera tener el grado de confiabilidad real, por lo general dicha población de muestra puede ser mayor al 15% de la población total de los datos. A continuación, podemos observar las actividades propuestas por la metodología en la fase evaluación.

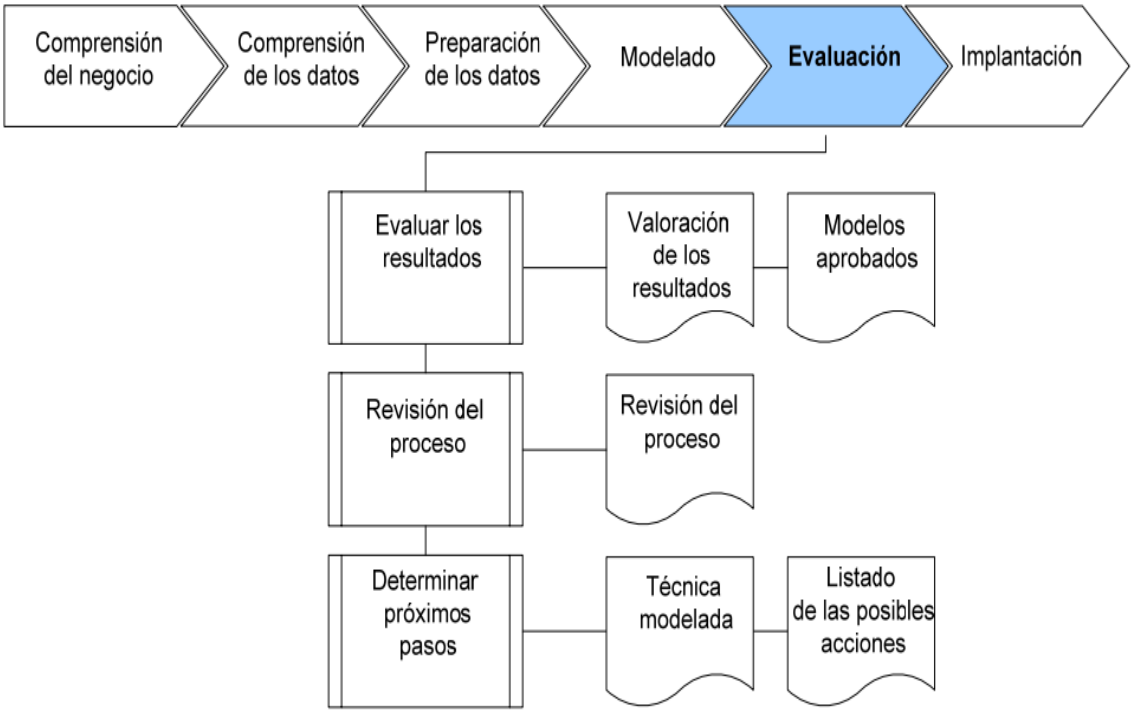


Ilustración 14. **Fase de Evaluación – Actividades**

Fuente. Obtenido de (José & Gallardo Arancibia, 2009)

- *Evaluar los Resultados:* En esta actividad se lleva a cabo la aplicación del proceso de contraste, Modelo Vs Realidad en la población excluida previamente al entrenamiento del

mismo, una vez se aplica el proceso correspondiente podemos obtener los positivos verdaderos, los positivos falsos, los negativos verdaderos y los negativos falsos adicionalmente del porcentaje de confiabilidad del modelo. El entregable en esta actividad puede ser una representación gráfica de la matriz de confusión.

- *Revisión del Modelo:* En esta actividad como su nombre lo indica se identifican aquellas mejoras o se puede dar la aceptación del modelo, según los criterios de aceptación definidos en las anteriores fases. El entregable es el análisis de la matriz de confusión Vs los criterios de aceptación del modelo.
- *Determinar próximos pasos:* En esta actividad se determina si existen ajustes y la forma de aplicarlos o sí definitivamente se puede seguir a la próxima fase.

Para el desarrollo del presente proyecto, se tiene planteado construir un modelo predictivo utilizando diferentes técnicas de análisis que posteriormente se contrastarán cada una de ellas para seleccionar la técnica que nos entregue el mejor nivel de confiabilidad para la ejecución de este proyecto.

3.1.6 Fase 6: Implementación

En esta fase, se lleva a cabo la conclusión del proyecto en donde se evidencia el conocimiento obtenido y respuesta a la gran pregunta formulada inicialmente por parte de la organización y el cual origino el desarrollo del proyecto de analítica de datos, en esta fase se debe documentar cada uno de los resultados de los análisis obtenidos en la ejecución del modelo. A continuación, podemos observar las actividades propuestas por la metodología en la fase implementación.

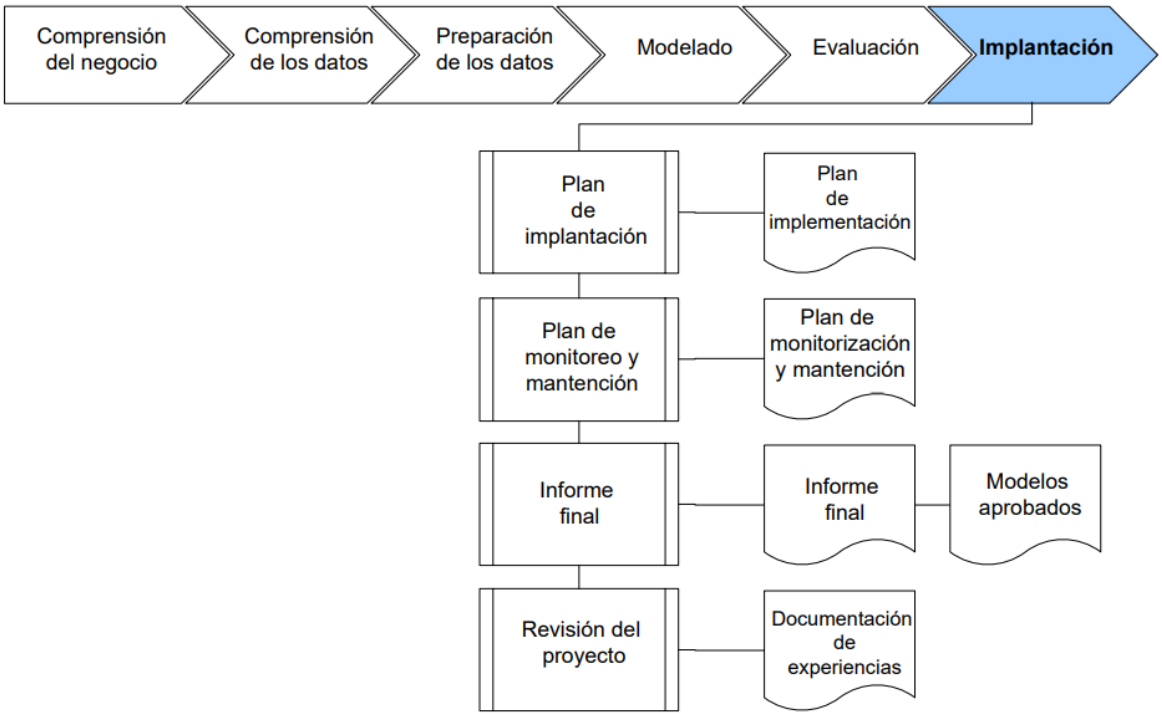


Ilustración 15. **Fase de Implementación – Actividades**

Fuente: Obtenido de (José & Gallardo Arancibia, 2009)

- *Plan de Implementación:* En esta actividad se establecen las actividades y se construye el plan de aplicación en la organización sobre el modelo desarrollado. El entregable de esta actividad es el plan de implementación.
- *Plan de monitoreo y mantención:* En esta actividad se deben establecer las estrategias de monitoreo y aplicación de insumos o recursos que necesite el modelo para su correcto funcionamiento, el entregable de esta actividad es el documento de estrategias y/o el plan de monitoreo propuesto.
- *Informe final:* En esta actividad se realizan las conclusiones del proyecto de analítica de datos desarrollado, en donde se establecen las mejores prácticas y puntos relevantes obtenidos en el desarrollo del proyecto.

- *Revisión del proyecto:* En esta actividad se evalúa sí existe o no una respuesta a la pregunta planteada inicialmente, las estrategias que se deben seguir en la organización para mejorar o mitigar el riesgo, según sea el caso.

Para la realización del presente proyecto de grado, se realizarán los análisis y conclusiones correspondientes sobre el conocimiento obtenido a través del respectivo procesamiento de los datos, y de esta manera presentar un análisis prescriptivo al agricultor, para que su cultivo tenga rendimientos más altos y adecuados, haciéndolo más rentable y por ende competitivo, mitigando al máximo posible los riesgos de su inversión al momento de cultivar Aguacate Hass en dicho territorio.

3.2 Herramientas

La herramienta propuesta para el desarrollo de este proyecto es la siguiente:

- **Microsoft Azure Machine Learning Studio.**

  Es un entorno interactivo de trabajo que funciona como una herramienta colaborativa de arrastrar y soltar. Se puede utilizar para crear, probar e implementar soluciones de análisis predictivo en sus datos (Ericson, Gilley, Martens & Anton, 2019). El entorno se encuentra totalmente en la “nube”, el acceso y operación es través de la web como se observa en la siguiente Ilustración:

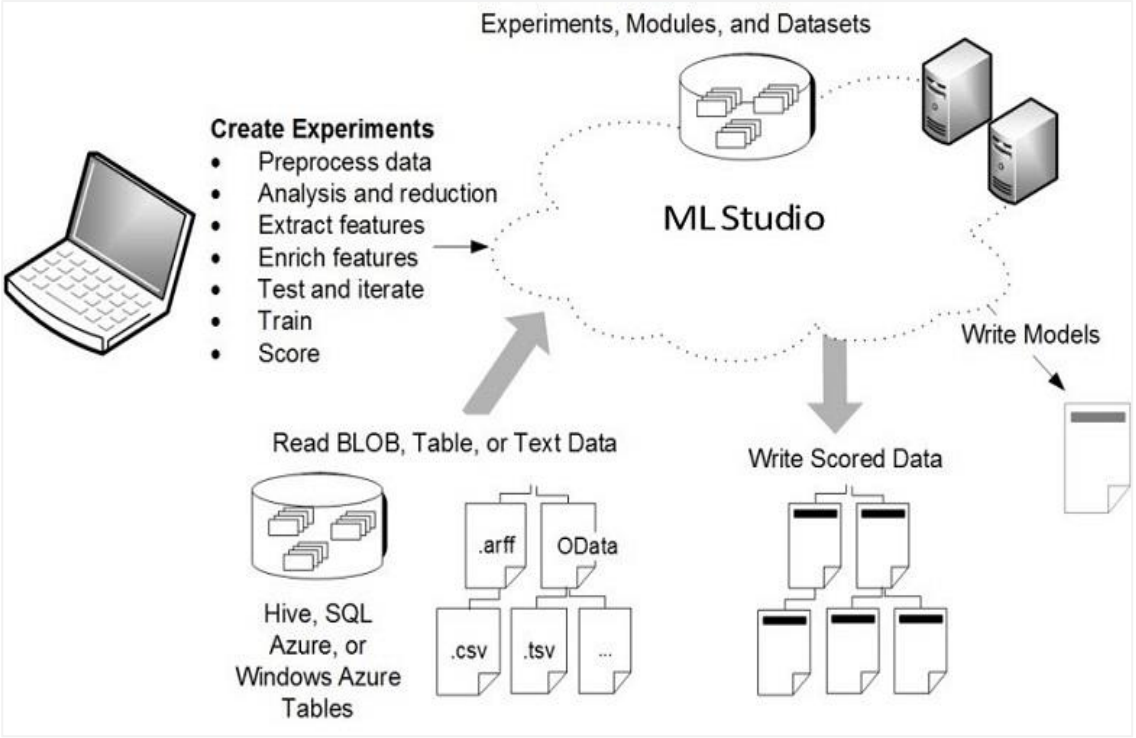


Ilustración 16. **Componentes del entorno de trabajo Microsoft Machine Learning Studio**

Fuente: Obtenido de (Ericson, Gilley, Martens & Anton, 2019).

- **Lenguaje R.**

  Los atributos por el cual los seleccionamos son: (Software Open Source, Fácil de Integrar con otras plataformas, Robusto en análisis estadístico, cuenta con bastante documentación).

CAPITULO 4. DESARROLLO

4.1 Comprensión del negocio

El principal objetivo del sector agrícola es ser cada vez más competitivo con respecto al mercado nacional e internacional. Por lo tanto, debe construir estrategias para generar acciones que se consoliden en una ventaja competitiva. En este contexto, la estrategia de la producción agrícola del aguacate Hass consiste en ser más eficientes, es decir, la estrategia está enfocada en incrementar el rendimiento de los cultivos con respecto a las cifras internacionales de los países más productores y con mayor exportación.

Para que un país alcance esta meta de manera sostenida debe iniciar su estrategia desde cada unidad productiva en referencia al pequeño y gran productor agrícola. Por lo tanto, surge la siguiente pregunta: ¿Cómo aumentar el índice de rendimiento en la producción del aguacate Hass en las organizaciones productoras agrícolas? Para resolver este interrogante, antes es necesario ahondar en los factores que influyen en es este indicador productivo, principalmente factores los externos como las variables meteorológicas, considerando que los factores internos se encuentran controlados bajo el estudio y empleo de buenas prácticas de producción.

4.1.1 Objetivos del negocio

Identificar los factores que determinarán las acciones que deben implementar una organización agrícola para aumentar el rendimiento en la producción de aguacate.

Obtener una herramienta analítica que permita predecir el rendimiento de la producción de aguacate a partir de datos meteorológicos actuales y futuros.

4.1.2 Valoración de la Situación

Colombia se encuentra en una posición favorable en cuanto a recursos naturales, condiciones medioambientales y ubicación geográfica. Según se establece, las condiciones climáticas y las cualidades heterogéneas de su geografía que la caracterizan, son las que favorecen o complican la eficiencia en la producción agrícola. Sin embargo, las acciones que se toman en la gestión de la producción agrícola van determinadas por la experiencia empírica y no por la medida en que estos factores influyen en su rendimiento.

Gracias al avance tecnológico y a las herramientas de análisis, como el aprendizaje de maquina (Machine Learning), cada vez más accesibles, es posible orientar las decisiones entorno a la información que nos brindan estas herramientas como un primer paso para las organizaciones, en particular las organizaciones productoras agrícolas, hacia la transformación digital.

Por lo tanto, se estima como hipótesis que: **El rendimiento del cultivo de aguacate Hass está condicionado por la influencia de las variables climáticas temperatura, precipitación y humedad relativa.**

4.1.3 Objetivos de la analítica de datos

Determinar las variables meteorológicas que más influyen en la producción del aguacate en las 18 regiones más productoras en Colombia

Obtener un modelo predictivo que clasifique el rendimiento de acuerdo a las variables climáticas proporcionadas.

4.1.4 Plan del proyecto

Recopilar información pertinente a la producción agrícola y sus factores influyentes

Organizar, explorar y estructurar la información

Elegir el indicador de éxito

Establecer un procedimiento de pruebas

Identificar un modelo adecuado y ajustarlo a los objetivos de la herramienta

Integrar la herramienta a los procesos de gestión de la organización.

4.2 Comprensión de los datos

4.2.1 Recolección de datos

Para iniciar la fase de recolección de datos se realizó una investigación por las distintas fuentes disponibles y asequibles en materia de información meteorológica y agrícola con suficientes registros históricos necesarios para el análisis del presente proyecto.

Tabla 6. ***Recolección de datos***

Variables	Tipo	Período	Fuente de información
Climáticas	Humedad relativa Temperatura Precipitación Brillo solar	Mensual	IDEAM
Producción agrícola	Área cosechada Área sembrada Producción Rendimiento	Mensual	AGRONET

Fuente: Elaboración Propia

Los datos meteorológicos fueron proporcionados por el IDEAM a través de una solicitud directa realizada por nosotros, que fue correspondida con el envío de un archivo plano(*.txt) con datos semiestructurados de periodicidad mensual por cada uno de los años (2007 – 2016), la información de las variables suministradas se relaciona en la tabla 7:

Tabla 7. ***Datos meteorológicos. IDEAM***

I D E A M - INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES

SISTEMA DE INFORMACION NACIONAL AMBIENTAL

VALORES MEDIOS MENSUALES DE HUMEDAD RELATIVA (%)

FECHA DE PROCESO : 2018/12/05 ESTACION : 23085080 NUS GJA EXP EL

LATITUD 0629 N TIPO EST CP DEPTO ANTIOQUIA FECHA-INSTALACION 1972-MAY

LONGITUD 7450 W ENTIDAD 01 IDEAM MUNICIPIO SAN ROQUE FECHA-SUSPENSION

ELEVACION 0835 m.s.n.m REGIONAL 01 ANTIOQUIA CORRIENTE NUS

A#O	EST	ENT	ENERO *	FEBRE *	MARZO *	ABRIL *	MAYO *	JUNIO *	JULIO *	AGOST *	SEPTI *	OCTUB *	NOVIE *	DICIE *	VR ANUAL *
2006	1	01	79 3	78 3	82 3	84 3	84 3	83 3	78 3	79 3	80 1	83 1	83 3	83 3	81 3
2007	1	01	82 3	74 3	81 3	*				83 3	83 3	86 3	86 3	85 3	83 3
2008	1	01	84 3	82 3	83 1	83 1	85 3	82 1	82 1	82 3	81 3	83 1	85 3	81 3	83 3
2009	1	01	81 3	82 3	84 1	84 3	82 1	82 1	83 1		80 3	84 1	85 3	83 3	83 3
2010	1	01	75 3	76 1	79	82	83 3	82 3	84	80 3	83 3	84 3	87 3	85 3	82 3
2011	1	01	80 3	80 3	83 3	85 3	84 3	82 3	81 1	78 3	79 3	84 3	*	*	82 3
2012	1	01	*	75 3	80 3	84 3	83 3	79 3	79 3	80 3	80 3	*		83 3	80 3
2013	1	01	76	78 3	79 3	*	83 3	81 3	78 1	81 3	78 3	79 3	82 3	82 3	80 3
2014	1	01	81 3				83 3	78 3	74	80 3	82 3	85	86 3	80 3	81 3
2015	1	01	79 3	81 3	80 3	*	83 3	79	76	81 3	79 3	*	90 1	89 3	82 3
2016	1	01	88	87	89 1	91 1	89 1	89	88 1	87 1	88 3				88 3
2017	1	01								79 3	83 1	84			82 3
MEDIOS			81	79	82	85	84	82	80	81	81	84	86	83	82
MAXIMOS			88	87	89	91	89	89	88	87	88	86	90	89	91
MINIMOS			75	74	79	82	82	78	74	78	78	79	82	80	74

Fuente: Obtenido de: IDEAM, 2018.

Para realizar la solicitud de los datos fue preciso estudiar los tipos de estaciones y la ubicación para posteriormente identificarlas en el listado nacional de estaciones publicado en el portal web del IDEAM y adicionalmente especificando el intervalo de tiempo en el cual se requería obtener los datos de las variables climáticas disponibles en las estaciones.

Los datos estadísticos de producción agrícola fueron descargados desde la página de AGRONET. AGRONET es una entidad encargada de centralizar y consolidar la información proporcionada por distintas instituciones, entidades, organizaciones y gremios para luego hacerlos accesibles en su portal. Los datos que disponen se encuentran organizados de forma estructurada como se muestra en la tabla 8.

Tabla 8. ***Datos estadísticos de producción***

Año	Municipio	Area Cos. (has)	Area Sem. (has)	Producción (Ton)	Rendimiento (ton/ha)
2007	Abejorral	83,00	96,00	705,50	8,50
2008	Abejorral	105,00	165,00	577,50	5,50
2009	Abejorral	105,00	165,00	577,50	5,50
2010	Abejorral	105,00	170,00	577,50	5,50
2011	Abejorral	104,00	239,00	208,00	2,00
2012	Abejorral	235,00	399,00	2.350,00	10,00
2013	Abejorral	276,00	376,00	3.864,00	14,00
2014	Abejorral	292,00	446,00	2.774,00	9,50
2015	Abejorral	317,00	436,00	3.011,50	9,50
2016	Abejorral	416,00	486,00	4.992,00	12,00

Fuente: Obtenido de AGRONET, 2017.

4.2.2 Descripción de los datos

4.2.2.1 Datos Meteorológicos.

Las estaciones climáticas están identificadas por un código único que relaciona todos los datos y características según el tipo, la ubicación y operación. Para un mismo código de estación existen distintos bloques de datos correspondientes con la variable climática registrada por su instrumental. El conjunto de datos (Dataset) proporcionado por el IDEAM, consiste de una tabla de 39738 filas y una columna, que contiene 1406 bloques de registros, cada uno con una cabecera de datos relacionados con 129 estaciones y 13 tipos de variables climáticas como se muestra en la tabla 9.

Tabla 9. ***Datos Meteorológicos***

Lista de variables suministradas	Unidad de medida
No días mensuales de precipitación	Milímetros (mms)
Totales mensuales de precipitación	Milímetros (mms)
Máximos mensuales de precipitación	Milímetros (mms)
Mínimos mensuales de temperatura	Grados Celsius (°c)
Máximos mensuales de temperatura	Grados Celsius (°c)
Medios mensuales de temperatura	Grados Celsius (°c)

Medios mensuales de tensión de Vapor	Milibares(mb)
Medios mensuales de punto de rocio	Grados Celsius (°c)
Totales mensuales de brillo Solar	Horas
Medios mensuales de humedad relativa	%
Medios mensuales de nubosidad	Octas
Totales mensuales de evaporación	Milímetros (mms)
Medios mensuales de velocidad del Viento	Metros por segundo (m/s)

Fuente: Elaboración propia.

4.2.2.2 Datos estadísticos agrícolas.

Los datos estadísticos de producción agrícola fueron seleccionados y descargados del sitio web oficial, suministrado por el Ministerio de Agricultura. Para el estudio del presente proyecto fueron seleccionados el top 18 de los departamentos con más producción de Aguacate Hass en Colombia en donde se involucraron las siguientes variables con periodicidad anual:

Tabla 10. ***Datos estadísticos agrícolas***

Lista de variables suministradas	Unidad de medida
Área cosechada	Hectáreas (has)
Área sembrada	Hectáreas (has)
Producción	Toneladas (ton)
Rendimiento	Ton/Has

Fuente: Obtenido de: https://www.agronet.gov.co

4.2.3 Exploración de los datos

A partir de los bloques de tablas del conjunto de datos se extrae una tabla representativa de los valores totales mensuales de la variable precipitación del municipio de Apartadó en el departamento de Antioquia, como se muestra a continuación:

Tabla 11. ***Valores totales mensuales de precipitación***

I D E A M - INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES

SISTEMA DE
NACIONAL

VALORES TOTALES MENSUALES DE PRECIPITACION (mms)

FECHA DE PROCESO : 2018/12/05 ESTACION : 12015020 UNIBAN

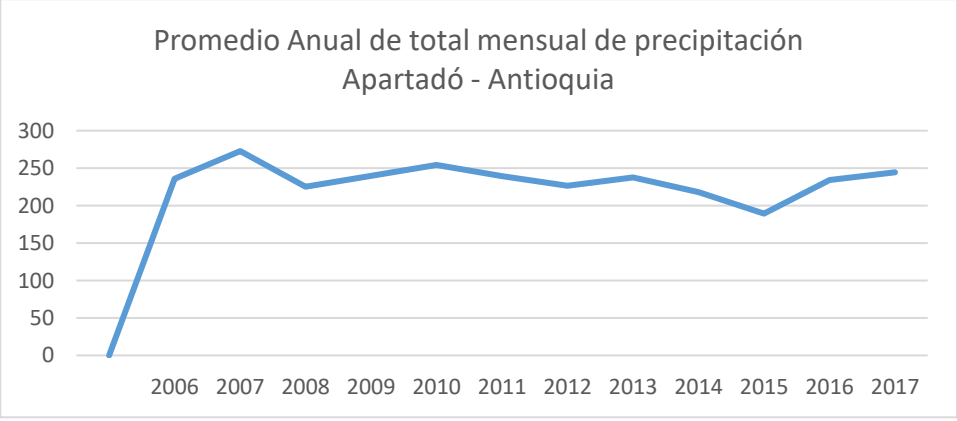
LATITUD	0749 N	TIPO EST	AM	DEPTO	ANTIOQUIA
LONGITUD	7639 W	ENTIDAD	01 IDEAM	MUNICIPIO	APARTADO
ELEVACION	0023 m.s.n.m	REGIONAL	01 ANTIOQUIA	CORRIENTE	ZUNGO

FECHA-INSTALACION
FECHA-SUSPENSION

A#O	EST	ENT	ENERO *	FEBRE *	MARZO *	ABRIL *	MAYO *	JUNIO *	JULIO *	AGOST *	SEPTI *	OCTUB *	NOVIE *	DICIE *
2006	2	01	118.9	94.8	188.3	338.1	576.6	110.3	199.5	240.9 3	151.1	406.8	242.5	161.6
2007	2	01	4.2	19.0	55.6	705.1	521.5	381.8	312.9	273.9	353.4	167.8	252.9	225.0
2008	2	01	86.6	250.8	63.7	175.0	355.6	335.3	269.7	220.8	234.3	341.5	238.5	131.7
2009	2	01	102.3	125.8	152.5	184.7	252.2	230.8	344.2	274.9	284.5	296.8	466.5	160.4
2010	2	01	.6	8.4	213.7	96.8	235.5	310.0	202.5	368.5	308.0	165.2	488.8	654.8
2011	2	01	259.3	117.4	173.3	398.1	252.7	128.3	385.9	303.0	228.6	179.3	220.2	223.9
2012	2	01	131.5	21.1	30.5	373.9	443.8	189.0	230.8	278.1	224.7	280.6	335.4	177.0
2013	2	01	.9	77.3	241.5	246.9	342.9 3	243.7	250.4	300.4	340.3	277.3	275.4	254.8
2014	1	01	96.9	18.6	71.3	103.9	647.2	205.2	276.3	218.0	184.8 3	268.2	288.2	237.2
2015	1	01	53.1	39.7	58.5	138.9 3	219.8	135.8	175.1	325.2	357.0	287.5	311.3	167.6
2016	1	01	1.0	20.2	15.1	332.7	192.3	300.3	290.0	130.9	278.8	308.8	448.7	492.8
2017	1	01	62.6	33.7	275.1	242.0	467.5	339.5	168.3	109.9	555.8	189.9		

Fuente: Obtenido de IDEAM.

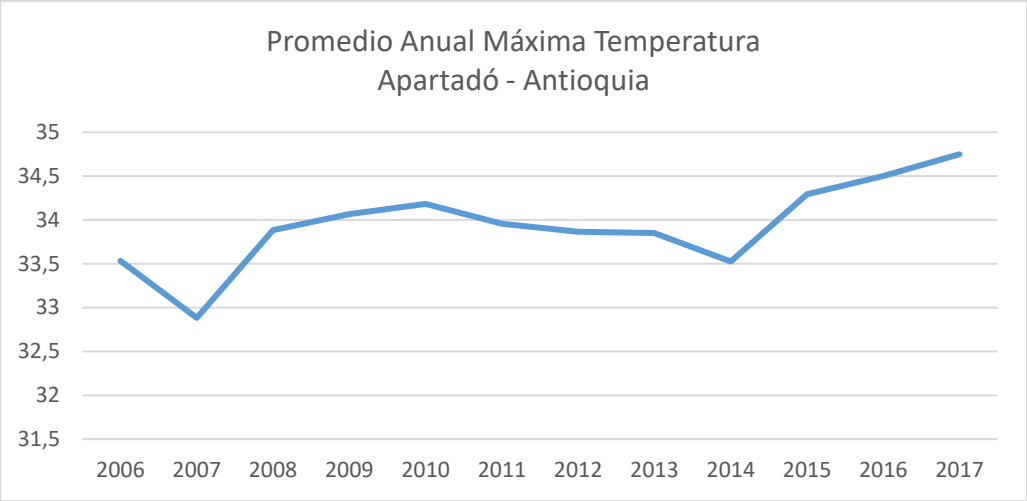
Se representa gráficamente el promedio de los valores mensuales de las variables, en este caso, el total mensual de precipitación del municipio de Apartadó Antioquia



Gráfica 9. **Promedio anual de total mensual de precipitación.**

Fuente. Obtenido de IDEAM

También son representados gráficamente los valores promedio anual de la máxima temperatura del municipio de Apartadó.



Gráfica 10. **Promedio anual máxima temperatura**

Fuente. Obtenido de IDEAM

Al aplicar un conjunto de técnicas de estadística descriptiva obtenemos los siguientes valores:

Tabla 12. ***Valores (Conjunto de estadísticas descriptivas)***

Año	Prom_Anual Max_Precipitacion	Prom_Anual Total_TEMP
Media	234,7406944	33,94116162
Error típico	5,838737873	0,141365093
Mediana	236,7166667	33,92083333
Moda	#N/A	#N/A
Desviación estándar	20,2259813	0,489703048
Varianza de la muestra	409,0903194	0,239809076
Curtosis	2,246598341	1,045613367
Coeficiente de asimetría	-0,492668096	-0,521025512
Rango	83,63333333	1,866666667
Mínimo	189,125	32,88333333
Máximo	272,7583333	34,75
Suma	2816,888333	407,2939394
Cuenta	12	12
Nivel de confianza (95,0%)	12,85097541	0,311142473

Fuente. Elaboración propia.

En las tablas de datos meteorológicos analizados es posible evidenciar una homogeneidad entre los distintos grupos de valores, pertinentes con las magnitudes representativas en el margen de valores reales de las escalas observadas y parametrizadas en los entornos naturales actuales. Sin embargo, se observa de manera frecuente la ausencia de registros de distintas variables en distintos periodos.

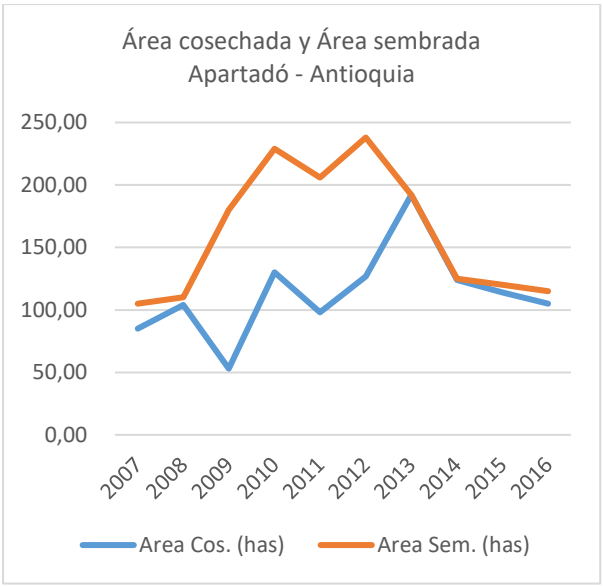
Para la exploración de los datos estadísticos agrícolas es extraída una muestra del conjunto de datos correspondiente al municipio de Apartadó en el departamento de Antioquia:

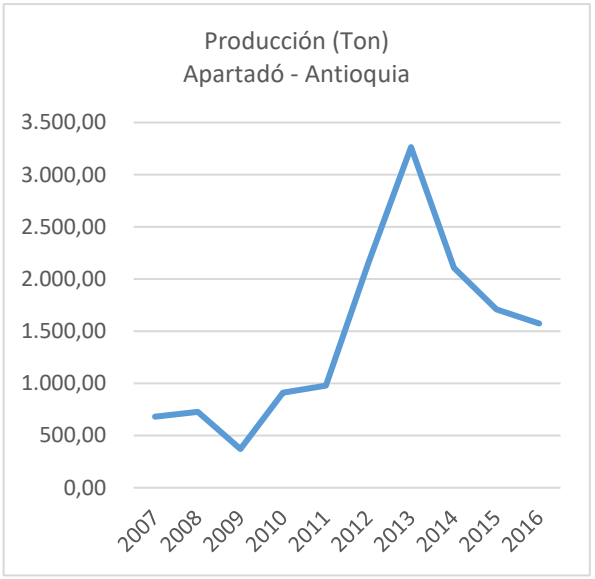
Tabla 13. C*onjunto de datos estadísticos Municipio de Apartadó*

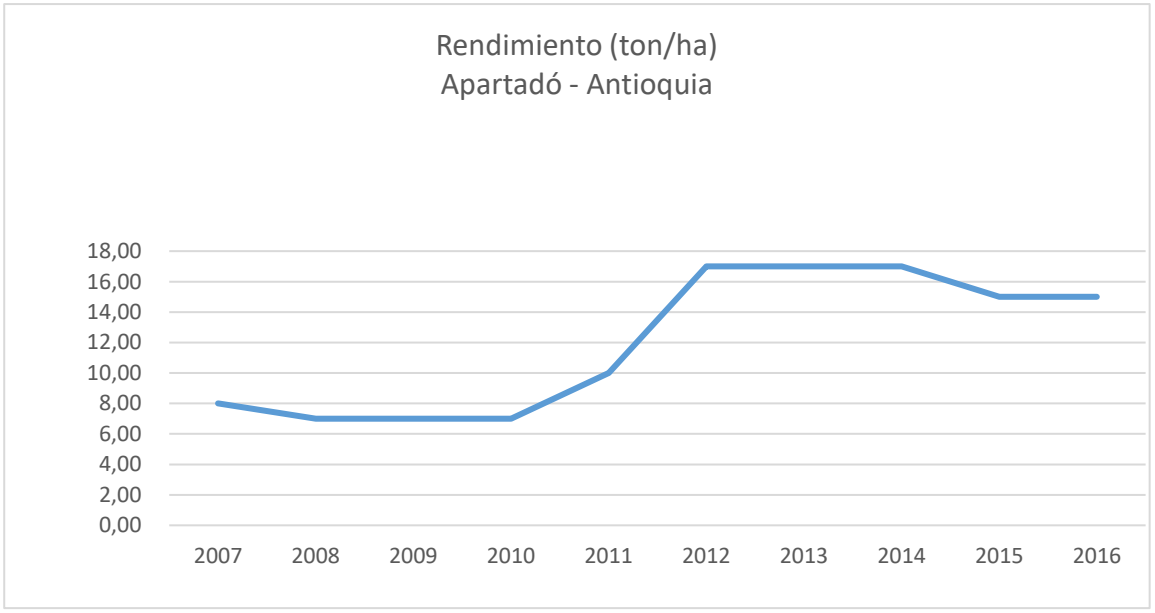
Año	Municipio	Area Cos. (has)	Area Sem. (has)	Producción (Ton)	Rendimiento (ton/ha)
2007	Apartadó	85,00	105,00	680,00	8,00
2008	Apartadó	104,00	110,00	728,00	7,00
2009	Apartadó	53,00	180,00	371,00	7,00
2010	Apartadó	130,00	229,00	910,00	7,00
2011	Apartadó	98,00	206,00	980,00	10,00
2012	Apartadó	126,80	238,00	2.155,60	17,00
2013	Apartadó	192,00	192,00	3.264,00	17,00
2014	Apartadó	124,00	125,00	2.108,00	17,00
2015	Apartadó	114,00	120,00	1.710,00	15,00
2016	Apartadó	105,00	115,00	1.575,00	15,00

Fuente. Obtenido de IDEAM

Al representar gráficamente cada variable agrícola se puede apreciar el comportamiento de las mismas en el intervalo de tiempo correspondiente.







Gráfica 11. **Comportamiento de variable agrícola del Municipio de Apartadó**

Fuente. Obtenido de IDEAM

Posteriormente se aplica un conjunto de técnicas de estadística descriptiva que permiten obtener una mejor perspectiva de la magnitud y comportamiento de los datos.

Tabla 14. ***Conjunto de Estadísticas descriptivas***

Estadística descriptiva	***Área Cos. (has)***	***Área Sem. (has)***	***Producción (Ton)***	***Rendimiento (ton/ha)***
Media	113,18	162	1448,16	12
Error típico	11,35102345	16,57307053	280,7835763	1,445298893
Mediana	109,5	152,5	1277,5	12,5
Moda	#N/A	#N/A	#N/A	7
Desviación estándar	35,89508787	52,40865069	887,9156306	4,5704364
Varianza de la muestra	1288,457333	2746,666667	788394,1671	20,88888889
Curtosis	2,56754593	-1,853912878	0,370264524	-2,246507793
Coeficiente de asimetría	0,76001802	0,29272501	0,855843954	-0,026185869
Rango	139	133	2893	10
Mínimo	53	105	371	7
Máximo	192	238	3264	17
Suma	1131,8	1620	14481,6	120
Cuenta	10	10	10	10
Nivel de confianza (95,0%)	25,677799	37,4908902	635,1765783	3,269493242

Fuente. Elaboración propia.

En la información estadística agrícola de los municipios de los 18 departamentos analizados, se observa la existencia de registros anuales a partir del año 2007 hasta el año 2016. Además, existe una deficiente recopilación de datos de los posibles municipios en las regiones productoras de cada departamento. Sin embargo, la información proporcionada es coherente en homogeneidad y magnitud con respecto a las características de las regiones.

4.2.4. Reporte de calidad de los datos

Tabla 15. ***Reporte – Entidad, clasificación y proceso de datos***

Entidad	**Clasificación**	**Proceso**
IDEAM	Preliminar IDEAM, preliminar otra entidad	Normalización con mínimos y máximos
AGRONET	Sin clasificación	Normalización con mínimos y máximos

Fuente. Elaboración propia.

4.3 Preparación de los datos

4.3.1 Seleccionar los datos

Para el desarrollo del presente proyecto fue necesario fue seleccionado la plataforma Microsoft Azure ya que tiene gran variedad de servicios que nos permiten ejecutar a cabalidad el desarrollo de los objetivos del presente estudio, los servicios adquiridos en dicha plataforma fueron:

- *SQL Database:* Se creó una base de datos transaccional de nombre Tesis para que nos permitiera la creación y el llenado de tablas estructuradas y de esta manera iniciar el análisis correspondiente.
- *Machine Learning:* Con este servicio nos permite generar el modelo predictivo necesario para cumplir con el objetivo del presente estudio, incluyendo la evolución y el Web Service necesario para las pruebas de dicho modelo.

En la siguiente pantalla se puede observar la estructura de la base de datos creada (Tesis).

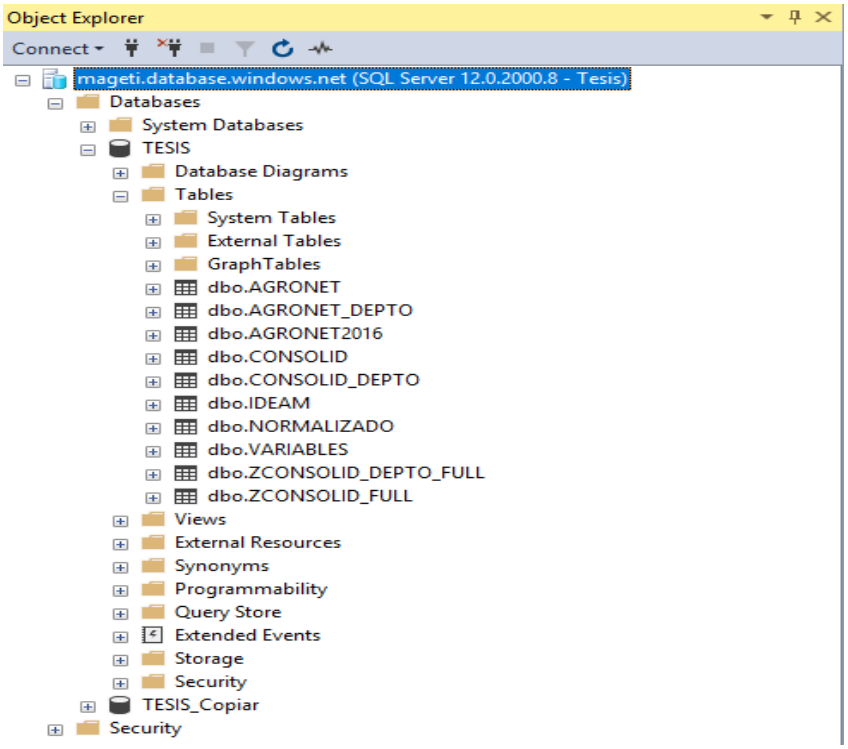


Ilustración 17. **Estructura de base de datos**

Fuente: Elaboración propia

Para consolidar la información suministrada por el IDEAM, fue necesario analizar, diseñar y desarrollar un desarrollo realizado en Visual Studio 2015 .Net con el lenguaje de desarrollo CSharp, que permitiera leer el archivo suministrado por la entidad y poblar la tabla “IDEAM”, creada en la base de datos (TESIS), en el desarrollo realizado se tuvieron en cuenta las siguientes reglas:

- No se tuvieron en cuenta valores vacíos.
- Solo se seleccionaron del encabezado del archivo suministrado los siguientes campos (Estación, tipo de estación, variable, municipio, departamento)

Los registros importados del archivo suministrado fueron: 90217 registros, como se observa en la siguiente imagen.

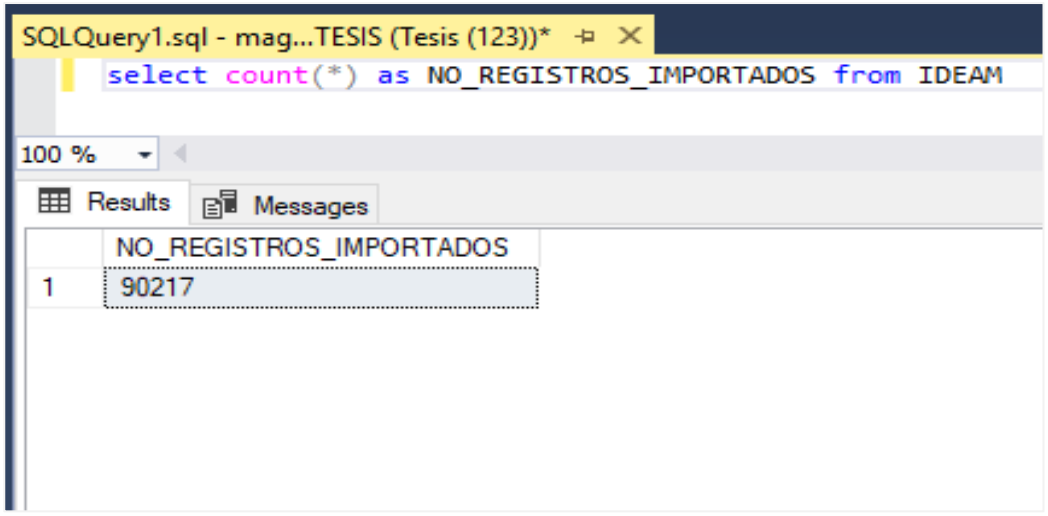


***Ilustración 18*. Registros importados de datos sin estructurar**

Fuente: Elaboración propia

Por otro lado, para la información seleccionada y descargada en el sitio web (Agronet), se realizó la importación de la información de una manera directa a nuestra base de datos en la tabla "AGRONET" ya que la información suministrada es estructurada, el resultado de esta fuente de información fueron 3322 registros, como se observa a continuación.

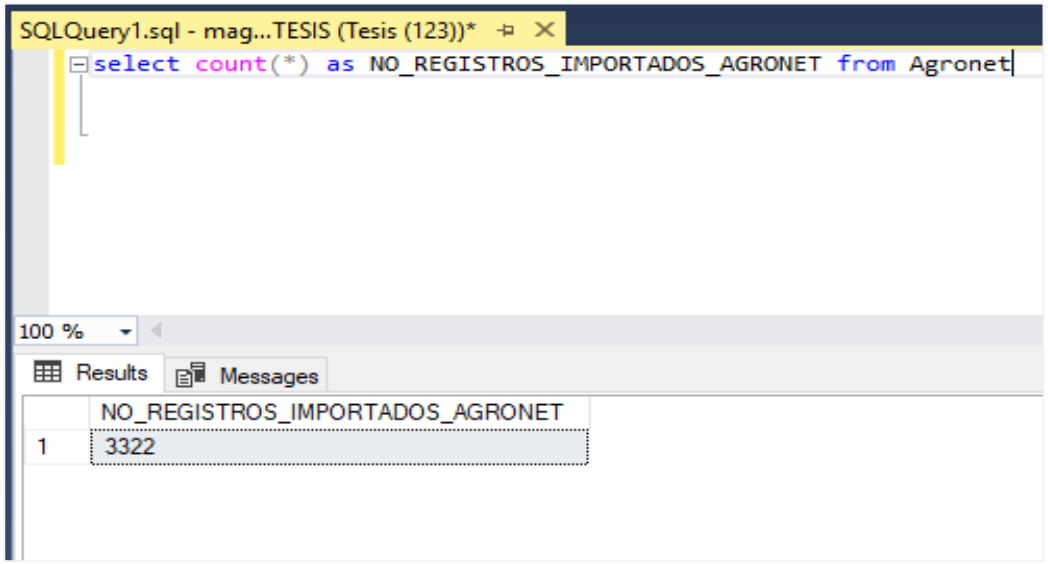


***Ilustración 19*. Registros importados de datos sin estructurar**

Fuente: Elaboración propia

El código CSharp generado para la ejecución de este proceso es entregado como anexo del presente estudio. Ejecutando el proceso anterior se lleva a cabo el proceso de consolidación de las dos fuentes de información en una base de datos estructurada.

4.3.2 Limpiar los datos

Con el propósito de hacer más eficiente el desarrollo del proceso, se hace énfasis en la depuración de la información contenida en el conjunto de datos meteorológicos. Son eliminados los registros sin información pertinente al tipo dato (numérico), registros duplicados y los registros con cantidades calculadas. Se elimina la información contenida fuera del período comprendido entre el año 2007 hasta el año 2016.

Para llevar a cabo este proceso y aprovechando las ventajas del motor de base de datos seleccionada se desarrolló un procedimiento almacenado y dos funciones escalares que permitieran la limpieza respectiva de los datos, con las siguientes reglas.

- Se depuran registros duplicados
- Se prioriza la información de las estaciones que contienen variables más precisas y completas de índole ambiental.
- Cálculo del promedio teniendo en cuenta solo los meses que tienen datos diferentes de 0 para que los promedios no fueran afectados y de esta manera poder generar registros promedio por año.

Los scripts desarrollados de los objetos anteriormente mencionados se adjuntan como anexo.

4.3.3 Estructurar datos

Una vez se ejecuta la limpieza de los datos y teniendo en cuenta que para el análisis se debe tener la información por departamento con una periodicidad anual, se aprovecharon los objetos desarrollados para crear y poblar una tabla (“CONSOLID”) de la base de datos previamente creada, teniendo como resultado la tabla pivoteada, a continuación, se realiza una muestra donde se puede evidenciar el resultado de los datos. Este proceso se realiza para la fuente de información IDEAM.

```
select * from CONSOLID
```

100 %

Results | Messages

	CON_ANIO	CON_MUNICIPIO	CON_DEPARTAMENTO	CON_T_EVAPORACION	CON_MIN_TEMP	CON_T_BSOLAR	CON_T_PRECIP	CON_MED_TVAPOR	CON_MED_TEMP	CON_MED_PROCIO	CON_MAX_TEMP	CON_MAX_PRECIP
238	2009	CONVENCION	NORTE DE SANTANDER	102.49	16.53	136.80	107.66	21.87	21.79	18.89	28.03	37.79
239	2010	CONVENCION	NORTE DE SANTANDER	110.82	16.68	143.65	204.57	22.48	22.26	19.34	29.31	45.17
240	2011	CONVENCION	NORTE DE SANTANDER	100.19	15.91	124.03	163.49	22.37	21.71	19.25	28.60	36.38
241	2012	CONVENCION	NORTE DE SANTANDER	113.59	15.55	147.56	128.58	22.23	22.00	19.15	29.18	33.97
242	2013	CONVENCION	NORTE DE SANTANDER	115.62	15.67	142.43	94.29	22.33	22.27	19.23	29.76	28.12
243	2014	CONVENCION	NORTE DE SANTANDER	106.59	15.52	176.64	120.91	22.17	22.76	19.08	30.62	37.53
244	2015	CONVENCION	NORTE DE SANTANDER	117.51	15.48	164.75	72.24	22.32	23.19	19.22	31.74	34.90
245	2007	CUCUTA	NORTE DE SANTANDER	176.93	20.40	188.33	71.58	25.19	27.28	21.11	35.93	25.62
246	2008	CUCUTA	NORTE DE SANTANDER	168.43	19.94	187.68	91.49	24.91	26.53	20.96	34.83	29.50
247	2009	CUCUTA	NORTE DE SANTANDER	170.28	20.68	186.74	74.87	24.88	27.04	20.90	34.73	28.19
248	2010	CUCUTA	NORTE DE SANTANDER	169.42	20.71	182.70	127.32	25.87	27.57	21.56	35.94	38.27
249	2011	CUCUTA	NORTE DE SANTANDER	149.68	20.39	168.93	129.59	25.25	26.46	21.15	34.89	42.10
250	2012	CUCUTA	NORTE DE SANTANDER	118.49	20.62	181.54	62.88	24.43	27.22	20.61	35.37	24.28
251	2013	CUCUTA	NORTE DE SANTANDER	172.61	21.00	181.14	63.56	24.68	27.55	20.75	36.19	30.47
252	2014	CUCUTA	NORTE DE SANTANDER	191.33	21.06	185.73	48.56	24.13	27.88	20.41	36.02	20.67
253	2015	CUCUTA	NORTE DE SANTANDER	157.37	19.22	138.77	42.74	23.55	29.03	20.03	36.62	14.57
254	2016	CUCUTA	NORTE DE SANTANDER	142.54	20.00	126.68	71.48	24.44	29.10	20.63	37.00	23.92
255	2017	CUCUTA	NORTE DE SANTANDER	119.43	19.40	121.04	57.14	24.71	28.11	20.80	34.24	21.89
256	2007	DUITAMA	BOYACA	71.53	5.57	103.89	164.95	12.67	12.20	10.31	19.40	28.35
257	2008	DUITAMA	BOYACA	66.90	5.15	141.68	189.39	12.07	11.55	9.61	19.27	17.74
258	2009	DUITAMA	BOYACA	67.53	4.25	116.31	84.78	10.38	9.68	7.37	19.53	18.98
259	2010	DUITAMA	BOYACA	64.09	4.23	137.43	134.25	10.54	10.20	7.61	20.42	25.40
260	2011	DUITAMA	BOYACA	61.39	4.20	77.86	151.59	10.49	9.64	7.57	19.97	24.24
261	2012	DUITAMA	BOYACA	65.78	7.55	102.26	141.23	11.64	11.92	9.05	19.40	25.78
262	2013	DUITAMA	BOYACA	67.52	7.55	106.21	152.55	12.00	12.19	9.50	19.92	29.91
263	2014	DUITAMA	BOYACA	63.28	3.43	112.72	106.71	10.56	10.38	7.67	20.05	23.78
264	2015	DUITAMA	BOYACA	64.36	7.13	117.71	83.88	12.35	12.32	9.95	19.80	20.38

Ilustración 20. **Datos estructurados del IDEAM**

Fuente: Elaboración propia

Por otro lado, para la información suministrada de la fuente AGRONET se realiza un script en la base de datos que permite la agrupación y cálculo de promedios por departamento y de esta manera estandarizar la información para que sea la línea base del análisis objetivo, como resultado de la ejecución de esta operación se obtiene la creación de la tabla “CONSOLID_DEPTO”, en nuestra base de datos estructurada. A continuación, se puede observar una muestra de la tabla mencionada.

```
select * from AGRONET_DEPTO
```

100 %

Results | Messages

	AGRO_DEPTO	AGRO_ANIO	AGRO_AREASEM	AGRO_AREACOS	AGRO_AREAPROD	RENDIMIENTO
1	CALDAS	2012	4637.20	2299.90	23679.20	10.30
2	CUNDINAMARCA	2010	392.60	272.30	1492.55	5.48
3	CHOCO	2013	218.00	276.00	752.50	2.73
4	BOLIVAR	2015	3561.00	2940.00	24135.00	8.21
5	BOYACA	2014	364.00	195.50	1296.00	6.63
6	META	2009	210.00	67.00	908.00	13.55
7	SUCRE	2009	616.00	269.00	1971.50	7.33
8	TOLIMA	2009	7349.50	4892.00	58201.65	11.90
9	HUILA	2012	775.80	295.80	2947.05	9.96
10	BOYACA	2008	81.50	70.50	813.70	11.54
11	SUCRE	2015	649.00	523.00	4103.25	7.85
12	QUINDIO	2010	980.30	716.50	4602.80	6.42
13	CESAR	2011	2082.50	1827.00	12938.00	7.08
14	LA GUAJIRA	2016	522.00	477.00	2258.00	4.73
15	QUINDIO	2007	668.40	453.80	3745.80	8.25
16	CAUCA	2014	639.50	425.30	2525.45	5.94
17	META	2015	780.00	527.00	5306.00	10.07
18	VALLE DEL CAUCA	2011	1772.50	1168.00	15636.30	13.39
19	ANTIOQUIA	2014	7416.40	4850.60	46599.63	9.61
20	RISARALDA	2012	1492.54	1260.95	12458.24	9.88
21	BOLIVAR	2007	4168.00	3308.00	43080.00	13.02
22	SANTANDER	2011	2438.20	2127.40	17799.00	8.37
23	CUNDINAMARCA	2007	195.26	130.90	1091.00	8.33
24	CHOCO	2016	284.50	312.50	1528.50	4.89
25	SUCRE	2012	633.50	600.00	4805.00	8.01
26	CALDAS	2009	1539.40	954.40	8326.50	8.72
27	NORTE DE SANTANDER	2011	171.50	152.50	1668.00	10.94

***Ilustración 21*. Datos estructurados de AGRONET**

Fuente: Elaboración propia

4.3.4 Integrar los datos

Una vez realizada la limpieza y la estructura de los datos, procedemos a crear la unión de las dos fuentes de información a través de un script generado en la base de datos, teniendo como resultado los datos en la misma escala y con la misma periodicidad, el cual nos permite tener la línea base de los datos para iniciar con el análisis respectivo para el cumplimiento de los objetivos del presente proyecto, a continuación, se puede observar el resultado obtenido.

Tabla 16. ***Datos combinados***

COND_T_PR	COND_MED_	COND_MED_	COND_MED_	COND_MAX_	COND_MAX_	COND_MED_	COND_MED_	COND_NDIA	COND_MED_	AGRO_DEPT	AGRO_ANIO	AGRO_AREA	AGRO_AREA	AGRO_AREA	RENDIMIENTO
265.98	24.61	23.69	20.50	31.00	55.76	5.26	1.86	18.62	83.87	ANTIOQUIA	2007	2925.90	1828.70	19984.10	10.93
268.29	24.55	23.59	20.46	30.59	54.46	5.43	1.80	19.35	83.26	ANTIOQUIA	2008	3847.90	2040.80	20102.80	9.85
217.25	24.65	24.16	20.51	31.32	53.85	5.15	1.84	17.55	81.55	ANTIOQUIA	2009	4895.80	2376.00	23115.00	9.73
274.28	24.68	24.10	20.52	31.54	54.92	5.41	1.71	19.93	82.10	ANTIOQUIA	2010	5626.80	2907.20	28819.25	9.91
275.73	24.37	23.72	20.26	31.02	56.83	5.33	1.53	20.00	82.08	ANTIOQUIA	2011	6279.60	3196.00	28316.33	8.86
220.01	24.38	24.18	20.23	31.47	48.90	5.36	1.94	17.93	80.01	ANTIOQUIA	2012	6927.30	4082.70	38055.60	9.32
225.34	24.84	24.40	20.54	31.88	52.19	5.34	2.05	18.37	80.59	ANTIOQUIA	2013	7173.70	4826.10	46584.00	9.65
207.93	24.41	24.22	20.18	31.43	50.36	4.99	0.65	16.33	80.07	ANTIOQUIA	2014	7416.40	4850.60	46599.63	9.61
166.37	23.92	24.73	19.85	32.39	46.70	4.70	2.36	13.43	78.83	ANTIOQUIA	2015	10889.80	5649.10	61690.45	10.92
203.40	24.41	24.14	20.21	32.04	49.69	5.04	3.29	16.43	80.69	ANTIOQUIA	2016	11321.20	5780.20	67031.92	11.60
171.52	28.38	26.38	23.10	33.93	43.70	5.00	0.86	15.73	83.42	ARAUCA	2012	302.00	148.00	2320.00	15.68
152.02	28.78	26.95	23.33	34.81	39.19	4.67	1.14	13.73	82.25	ARAUCA	2013	309.00	160.00	2500.00	15.63
149.29	28.08	27.02	22.90	34.58	35.90	4.75	0.65	14.00	80.25	ARAUCA	2014	311.00	194.00	3010.00	15.52
151.05	28.77	26.88	23.33	34.41	44.80	4.50	2.36	13.42	82.50	ARAUCA	2015	316.00	258.00	4020.00	15.58
160.32	29.79	27.17	23.92	34.97	40.16	5.18	3.29	13.60	84.36	ARAUCA	2016	272.00	240.00	3750.00	15.63
145.31	30.44	27.75	24.27	35.30	50.32	3.71	1.68	10.55	82.09	BOLIVAR	2007	4168.00	3308.00	43080.00	13.02
121.74	30.14	27.74	24.13	34.33	39.94	4.19	1.67	10.25	81.35	BOLIVAR	2008	4378.00	3475.00	45180.00	13.00
88.60	30.12	28.24	24.12	34.87	34.89	3.49	2.24	8.27	79.46	BOLIVAR	2009	3983.00	3890.00	39992.00	10.28
189.10	31.00	28.16	24.60	34.36	53.88	4.07	2.92	13.29	81.76	BOLIVAR	2010	3983.00	3533.00	35304.00	9.99
162.04	30.08	27.79	24.08	34.30	48.65	3.96	2.74	12.22	81.20	BOLIVAR	2011	3983.00	3493.00	34990.00	10.02
89.69	29.08	28.30	23.48	34.77	33.09	3.52	1.98	8.81	77.00	BOLIVAR	2012	3771.00	3406.00	34804.00	10.22
114.98	30.04	28.42	24.06	35.06	37.76	3.48	2.29	10.15	78.72	BOLIVAR	2013	3967.00	2886.00	30248.00	10.48

4.3.5 Formateo de los datos

Para facilitar la generación del modelo se excluyen del conjunto de datos las columnas con variables categóricas no representativas y se discretiza el parámetro *Rendimiento* para facilitar la interpretación de los resultados en la etapa posterior de evaluación del modelo.

La discretización del parámetro rendimiento se realiza con base en la tabla 17:

Tabla 17. ***Parámetros de discretización de la variable rendimiento***

Rango de valores	**Categoría**
4 > x >10	BAJO
10 > x >14.49	ACEPTABLE
x >14.5	BUENO

4.3 Modelado

4.4.1 Técnica de modelado

La técnica apropiada para el desarrollo del modelo es un método que desempeñe tareas de regresión como de clasificación para identificar las variables que más influyen en el objetivo del negocio, así como para obtener un modelo predictivo en función de los parámetros climáticos. Las técnicas del modelo que más se ajustan con los requerimientos son Chi-cuadrado, la Regresión Logística Multiclase y Random Forest Multiclase. Los resultados de la regresión logística y el bosque aleatorio son comparables para conjuntos de datos más pequeños con menos de 1000 observaciones(Kirasich, Smith, & Sadler, 2018). Por lo tanto, se analizará los resultados con los dos métodos.

4.4.2 Generar el plan de pruebas

Para medir el progreso del modelo se utilizará el método de validación “hold out” y el método de validación “k-fold”. Ambos métodos serán aplicados para analizar la robustez del modelo en comparación del conjunto de pruebas y su aplicación.

Tabla 18. ***Plan de pruebas***

Método	**Descripción**	**Resultados calificados**	**Evaluación de resultados**
Hold-out	Divide el Dataset en dos partes: una grande para entrenamiento y otra para pruebas	Matriz de confusión	Métricas

K-Folds	Divide el Dataset en partes relativamente iguales y realiza el entrenamiento con las partes restante y lo evalúa con la actual	Matriz de confusión	Métricas

4.4.3 Construir el modelo

Para corroborar o rechazar la hipótesis analizamos la dependencia de las variables con respecto a la variable ESTADO, que es el resultado de discretizar la variable rendimiento.

1. Importamos los datos de entrada desde la base de datos (TESIS), teniendo en cuenta que está creada en la misma plataforma entonces nos permite crear una conexión directa al Dataset.

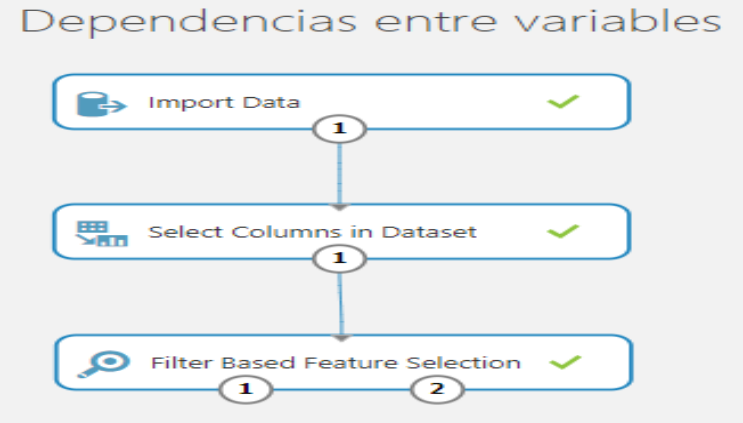


Ilustración 22. **Estructura de bloques para importar, seleccionar y filtrar datos**

Fuente: Elaboración propia

2. Seleccionamos las columnas con valores nominales a excepción de la variable ESTADO

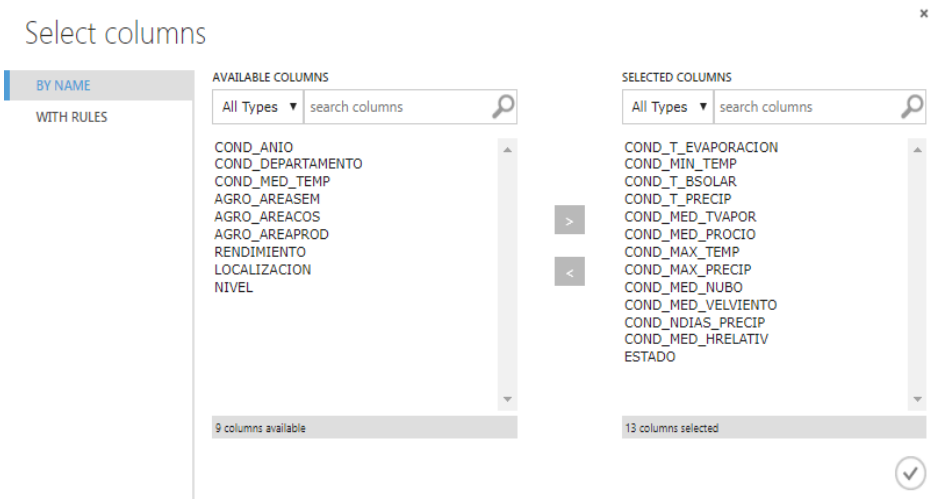


Ilustración 23. **Selección de columnas para correlacionar**

Fuente: Elaboración propia

3. **Escogemos como método de filtro el Chi Cuadrado para la variable ESTADO con un número de 5 características deseadas.**

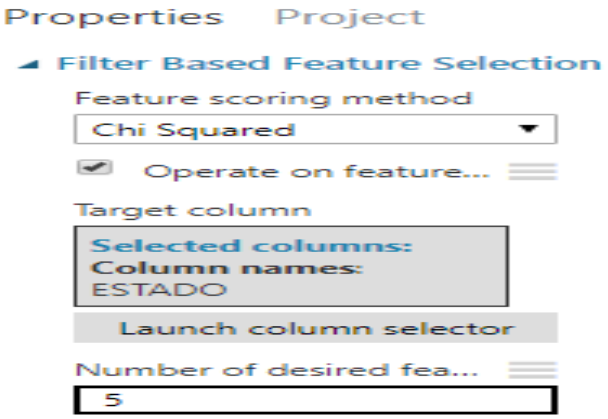


Ilustración 24. **Selección del método**

Fuente: Elaboración propia

4. **El resultado es el siguiente:**

   Dataset Filtrado

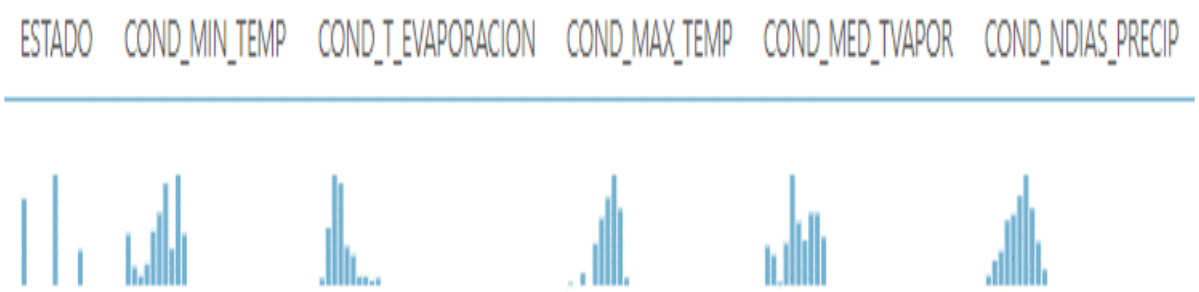


***Ilustración 25*. Resultado gráfico de la correlación**

Fuente: Elaboración propia

Características

ESTADO	COND_MIN_TEMP	COND_T_EVAPORACION	COND_MAX_TEMP	COND_MED_TVAPOR	COND_NDIAS_PRECIP
1	65.146841	59.276909	56.806021	46.898142	43.479447

***Ilustración 26*. Valores de los parámetros de la correlación**

Fuente: Elaboración propia

Para cual se puede deducir que las variables de mayor impacto son la temperatura en su valor mínimo y máximo, el total de evaporación, la media de tensión del vapor y el número de días de precipitación.

Dentro del mismo proceso se exportan automáticamente el resultado de los datos normalizados en la tabla “NORMALIZADO”, de la base de datos (TESIS) para que posteriormente

se pueda integrar a la herramienta Power BI de Microsoft para su respectivo análisis. A continuación, se puede ver una muestra de la tabla normalizada.

```
select * from NORMALIZADO
```

100 %

Results | Messages

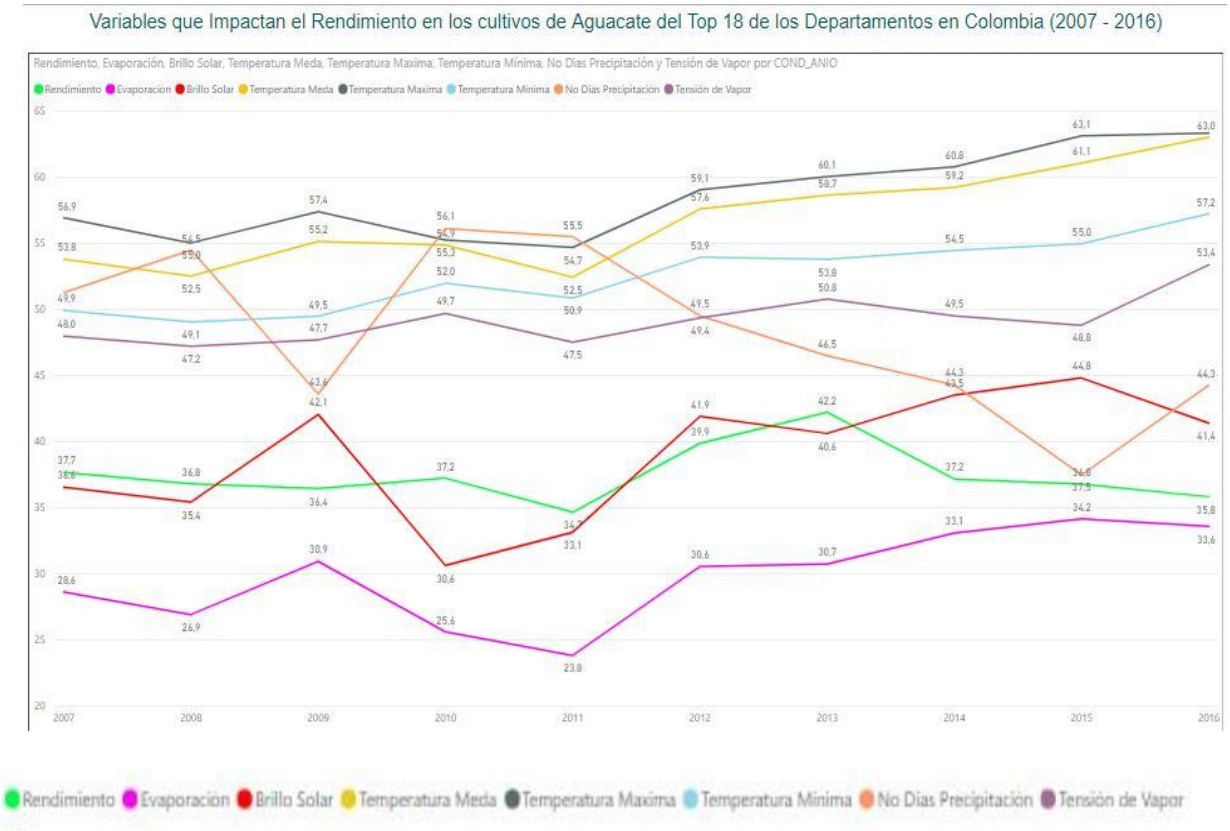
	id	COND_ANIO	COND_DEPARTAMENTO	COND_MIN_TEMP	COND_MAX_TEMP	COND_T_EVAPORACION	COND_T_BSOLAR	COND_MED_TVAPOR	COND_MED_PROCIO	RENDIMIENTO	COND_MED_VELVIENTO	COND_NDIAS_
1	1	2007	ANTIOQUIA	0,6442989634479	0,6602585349685 12	0,2371798846654 68	0,4270927092709 27	0,6635610766045 55	0,7431592039800 99	0,5855148342059 34	0,1154580152671 76	0,7392147358
2	2	2007	BOLIVAR	0,8979814511729 4	0,8027842227378 19	0,4913499100886 71	0,5682782563970 68	0,9653209109730 85	0,9776119402985 08	0,7678883071553 23	0,0982824427480916	0,3480368395
3	3	2007	BOYACA	0,0589198036006547	0,4126615843553 2	0,1973088609164 75	0,2900861514722 9	0,1330227743271 22	0,2220149253731 34	0,4109947643979 06	0,0982824427480916	0,6311197285
4	4	2007	CALDAS	0,0278232405891981	0,4149817693701 69	0,1787065170211 45	0,0715571557155 715	0,0087991718426501 9	0,0080845771144278	0,3089005235602 09	0,0667938931297 71	0,8444013572
5	5	2007	CAUCA	0,4266230223677 03	0,5926419705005	0,1622744465802 69	0,1713385624276 71	0,4570393374741 2	0,5267412935323 38	0,0872600349040 14	0,0944656488549 618	0,7421231216
6	6	2007	CESAR	0,8696126568466 99	0,8432217434537 62	0,5130526446332 24	0,8008872315803 01	0,8405797101449 28	0,8886815920398 01	0,2940663176265 27	0,1402671755725 19	0,4590402326
7	7	2007	CUNDINAMARCA	0,2913256955810 15	0,5674511103745 44	0,2435046815898 8	0,4105696283914 1	0,3436853000270 39	0,4179104477611 94	0,3586387434555 97	0,1354961832061 07	0,6107610276
8	8	2007	HUILA	0,6830332787779 96	0,7616837918462 05	0,3279593228746 82	0,2610261026102 61	0,5305330227743 3	0,6374378109452 74	0,6858638743455 5	0,1001908396946 56	0,5603490063
9	9	2007	LA GUAJIRA	0,9067103109656 3	0,8120649651972 16	0,6910770757115 4	0,8640864086408 64	0,9068329813664 6	0,9365671641791 04	0,0104712041884817	0,2709923664122 14	0,2544837615
10	10	2007	META	0,7670485542825 97	0,7507457739476 3	0,3440813542506 36	0,4432943293429 43	0,7179089026915 11	0,8009950248756 22	0,8699825479930 19	0,0696564885496 183	0,7576345128
11	11	2007	NORTE DE SANTANDER	0,3824331696672 12	0,5581703679151 48	0,3213244868853 47	0,4362221936479 36	0,4223602484472 05	0,5335820895522 39	0,7408376963350 79	0,1192748091603 05	0,5070285991
12	12	2007	QUINDIO	0,5471903982542 28	0,6297646668876 37	0,3109691821169 47	0,4276713385624 28	0,4534161490683 23	0,5827114427860 7	0,3516579406631 76	0,0677480916030534	0,7901114881
13	13	2007	RISARALDA	0,5821058374249 86	0,6025853496851 18	0,2553481738699 08	0,5171017101710 17	0,4057971014492 75	0,5366915422885 57	0,4415357766143 11	0,0992366412213 74	0,9355307804
14	14	2007	SANTANDER	0,5673758865248 23	0,6284388465362 94	0,3190922056179 08	0,5479619390510 48	0,5507246376811 59	0,6175373134328 36	0,4031413612565 44	0,1230916030534 35	0,6039747939
15	15	2007	SUCRE	0,8363338788870 7	0,7978123964202 85	0,3035902523717 99	0,1301915905876 3	0,8680124223602 48	0,9135572139303 48	0,1989528795811 52	0,0868320610687023	0,4178380998
16	16	2007	TOLIMA	0,8161483906164 76	0,8548226715280 08	0,5004650585973 83	0,6013887102996 01	0,7194616977225 67	0,8022388059701 49	0,4371727748691 1	0,1364503816793 89	0,4149297140
17	17	2007	VALLE DEL CAUCA	0,6808510638297 87	0,6957242293669 21	0,3284553853785 58	0,3424195562813 42	0,6097308488612 84	0,7108208955223 88	0,5846422338568 94	0,0391221374045802	0,6383906931
18	18	2008	ANTIOQUIA	0,6442989634479	0,6466688763672 52	0,2340794940162 46	0,3795165230808 8	0,6604554865424 43	0,7406716417910 45	0,4912739965095 99	0,1097328244274 81	0,7746000969
19	19	2008	BOLIVAR	0,8810692853246 04	0,7706330792177 66	0,4463322378619 71	0,6309630963096 31	0,9479729606625 26	0,9689054726368 16	0,7661431064572 43	0,0973282442748092	0,3334949103
20	20	2008	BOYACA	0,0861974904528096	0,4083526682134 57	0,1831090717430 4	0,3372765848013 37	0,0983436853002 071	0,1759950248756 22	0,6387434554973 82	0,1450381679389 31	0,6408143480
21	21	2008	CALDAS	0,0136388434260775	0,4070268478621 15	0,1714516029019 66	0,0276456217050276	0,0056935817805 3836	0,0037313432835 8201	0,3333333333333 33	0,1450381679389 31	0,9418322830
22	22	2008	CAUCA	0,4315330060010 91	0,5667882001988 73	0,1539653996403 55	0,1208049376366 21	0,4658385093167 7	0,5317164179104 48	0,0340314136125655	0,1106870229007 63	0,8618516723
23	23	2008	CESAR	0,8499727223131 48	0,8369240967848 86	0,5612947231351 15	0,7652050919377 65	0,8141821946169 77	0,8694029850746 27	0,1867364746945 9	0,1383587786259 54	0,4498303441
24	24	2008	CUNDINAMARCA	0,2613202400436 44	0,5303281405369 57	0,2028895640850 75	0,3921820753503 92	0,3540372670807 45	0,4421641791044 78	0,5767888307155 32	0,1116412213740 46	0,6892874454
25	25	2008	HUILA	0,6737588652482 27	0,7199204507789 19	0,2998697835927 33	0,2394239423942 39	0,5434782608695 65	0,6523631840796 02	0,6623036649214 66	0,1183206106870 23	0,6655356277
26	26	2008	LA GUAJIRA	0,8810692853246 04	0,7988067616837 92	0,7139579587027 96	0,8783592644978 78	0,8581780538302 28	0,9023631840796 02	0,0148342059336824	0,2232824427480 92	0,2287930198
27	27	2008	META	0,7670485542825 97	0,7421279416639 05	0,3419730886091 65	0,3533496206763 53	0,7013457556935 82	0,7885572139303 48	0,5191972076788 83	0,0648549616320 61	0,6505089675

***Ilustración 27*. Valores de los parámetros de la correlación**

Fuente: Elaboración propia

Una vez se obtienen estos datos se utilizó la herramienta de Power BI de Microsoft, en donde se creó una conexión directa a la tabla “NORMALIZADOS” de base de datos (TESIS) de SQL Database, creando una gráfica y así analizar el comportamiento de las variables normalizadas y con el mayor impacto identificadas por el método Chi Cuadrado, con la ventaja de actualizarse de manera automática y en línea ya que todos los componentes y/o servicios se encuentran bajo el

mismo ambiente para nuestro caso Azure. A continuación, se puede observar la representación gráfica en Power BI.



***Gráfica 12*. Gráfica de variables climatológicas y el rendimiento**

Fuente: Elaboración propia

4.4.3.1 Modelo predictivo

Modelo de regresión usando la regresión logística multiclase

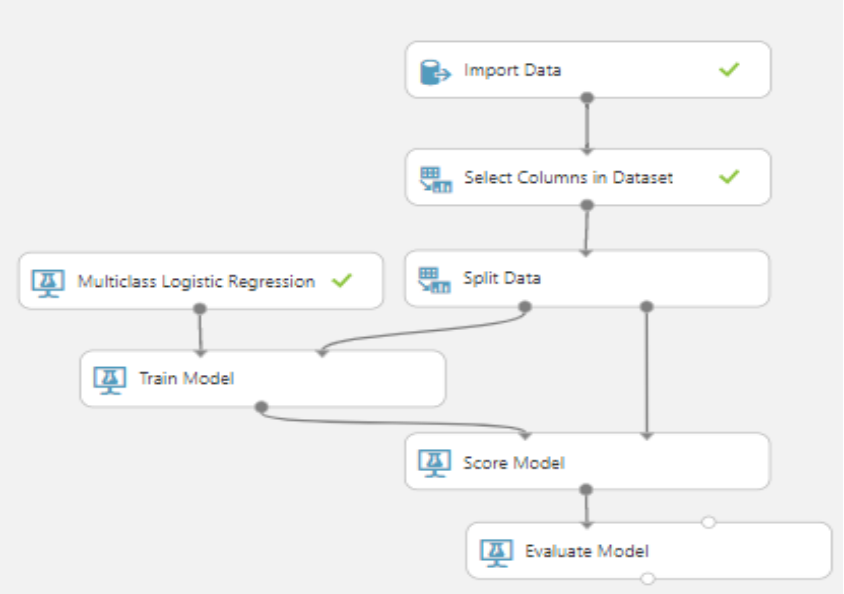


***Ilustración 28*. Estructura de bloques del Modelo Predictivo**

Fuente: Elaboración propia

1. Importamos el conjunto de datos (Dataset) de la base de datos (TESIS) y seleccionamos las columnas con valores nominales a excepción de la variable ESTADO.

2. Empleando el método validación en espera (*Hold out*) dividimos los datos en dos partes; una parte para entrenamiento del modelo con el 90% y la otra parte para la prueba con el 10% de los datos.

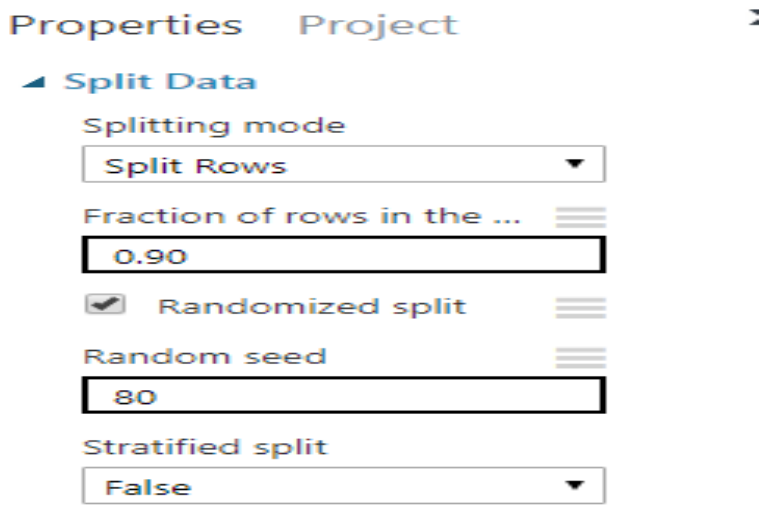


***Ilustración 29*. Datos para entrenamiento y prueba del modelo**

Fuente: Elaboración propia

3. Agregamos el bloque con el método de entrenamiento del modelo, en este caso la regresión logística multiclase.

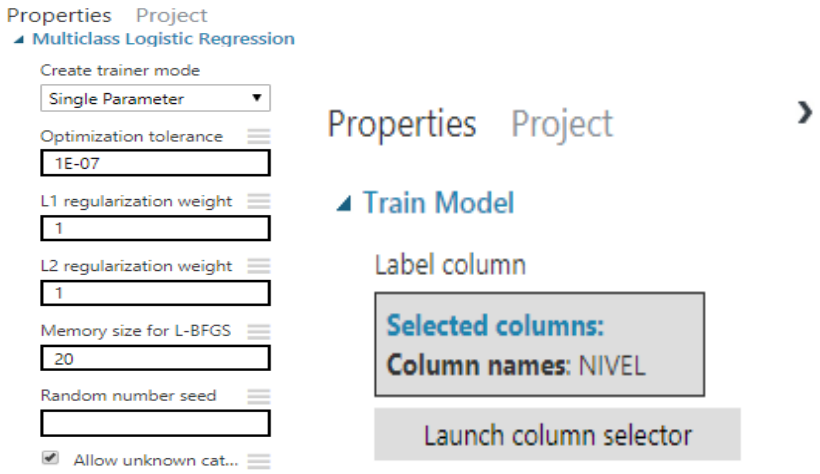


***Ilustración 30*. Método del entrenamiento del modelo**

Fuente: Elaboración propia

4. Los resultados se muestran a continuación

NIVEL	Scored Probabilities for Class "ACEPTABLE"	Scored Probabilities for Class "BAJO"	Scored Probabilities for Class "BUENO"	Scored Labels
BAJO	0.307207	0.512698	0.180095	BAJO
ACEPTABLE	0.54799	0.287692	0.164318	ACEPTABLE
BAJO	0.24713	0.588738	0.164132	BAJO
ACEPTABLE	0.55813	0.273489	0.168381	ACEPTABLE
ACEPTABLE	0.313769	0.499942	0.18629	BAJO
ACEPTABLE	0.481871	0.387162	0.130966	ACEPTABLE
ACEPTABLE	0.473165	0.369406	0.157428	ACEPTABLE
ACEPTABLE	0.585013	0.262256	0.152731	ACEPTABLE
BAJO	0.476541	0.383974	0.139485	ACEPTABLE

***Ilustración 31*. Resultados del modelo**

Fuente: Elaboración propia

Modelo de regresión usando el algoritmo de Bosque de decisión (resultados)

NIVEL	Scored Label Mean	Scored Label Standard Deviation
ACEPTABLE	0.34375	0.71739
BUENO	0.770833	0.873759
ACEPTABLE	0.8125	0.658478
BAJO	0.25	0.559017
BAJO	1.125	0.73951
BAJO	0.833333	0.57735
ACEPTABLE	0.375	0.484123
BUENO	0.5	0.866025
BAJO	0.666667	0.485913

***Ilustración 32*. Resultados del algoritmo Bosque de decisión**

Fuente: Elaboración propia

Si se emplea el modelo de validación cruzada con el método K-fold obtenemos los siguientes resultados.

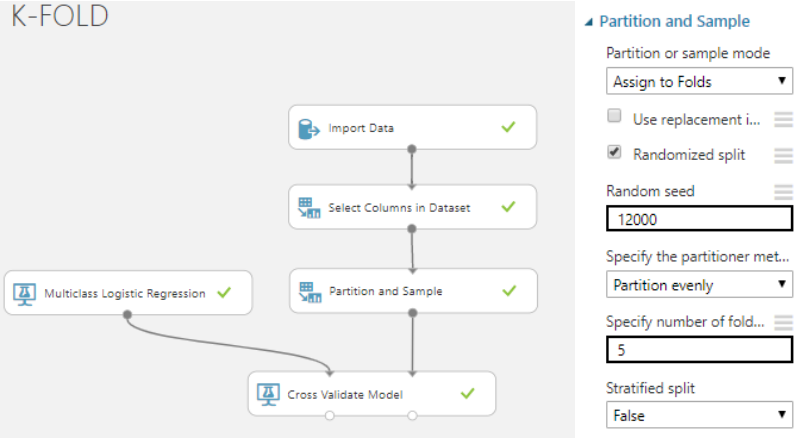


***Ilustración 33*. Validación cruzada con método K-Fold**

Fuente: Elaboración propia

Para aplicar la validación cruzada con el método K-Fold en Microsoft ML Studio, se reemplaza el bloque de Entrenamiento del Modelo (Train Model) y el bloque Evaluar Modelo (evaluate Model) por el bloque de Modelo de Validación Cruzada. En lugar de utilizar el bloque de Modo de Partición, se reemplaza por el bloque de Muestra y Partición con la configuración de particiones en "asignar a pliegues" (Assign to Fold) acompañada de una semilla aleatoria.

Regresión logística multilineal (K-Fold)

Al aplicar el método de regresión logística multilineal como método de entrenamiento obtenemos los siguientes resultados:

NIVEL	Scored Probabilities for Class "ACEPTABLE"	Scored Probabilities for Class "BAJO"	Scored Probabilities for Class "BUENO"	Scored Labels
ACEPTABLE	0.549937	0.263089	0.186974	ACEPTABLE
BUENO	0.440258	0.411916	0.147826	ACEPTABLE
ACEPTABLE	0.4869	0.407549	0.105551	ACEPTABLE
BAJO	0.693545	0.243922	0.062533	ACEPTABLE
BAJO	0.564898	0.272421	0.162681	ACEPTABLE
BAJO	0.40663	0.410725	0.182645	BAJO
ACEPTABLE	0.519513	0.359158	0.121328	ACEPTABLE
BUENO	0.49203	0.371847	0.136123	ACEPTABLE
BAJO	0.298535	0.528809	0.172656	BAJO

***Ilustración 34*. Resultados regresión logística multilineal método K-Fold**

Fuente: Elaboración propia

Si se aplica el método de Bosque de Decisión Multilineal en el modelo de validación cruzada se obtiene los siguientes resultados:

NIVEL	Scored Probabilities for Class "ACEPTABLE"	Scored Probabilities for Class "BAJO"	Scored Probabilities for Class "BUENO"	Scored Labels
ACEPTABLE	1	0	0	ACEPTABLE
BUENO	0.625	0.125	0.25	ACEPTABLE
ACEPTABLE	0.375	0.625	0	BAJO
BAJO	0.75	0.25	0	ACEPTABLE
BAJO	0	1	0	BAJO
BAJO	0	1	0	BAJO
ACEPTABLE	0.375	0.625	0	BAJO
BUENO	0.75	0	0.25	ACEPTABLE
BAJO	0.25	0.75	0	BAJO

***Ilustración 35*. Resultados método Bosque de Decisión Multilineal**

Fuente: Elaboración propia

Evaluación del modelo

La evaluación del modelo se hará en la siguiente fase.

4.5 Evaluación

4.5.1 Evaluación de los Resultados

Método Validación en espera (*Hold-Out)* & Regresión Logística Multilineal

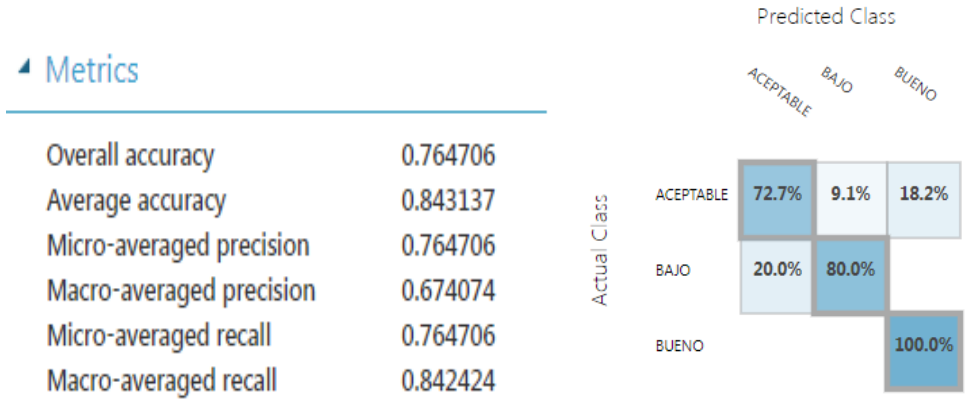


Ilustración 36. **Matriz de confusión y métricas del modelo empleando el método Hold-out y la regresión logística multilineal**

Fuente: Elaboración propia basada en los resultados de la plataforma de Microsoft ML Studio

Método Hold-Out & Bosque de Decisión Multilineal

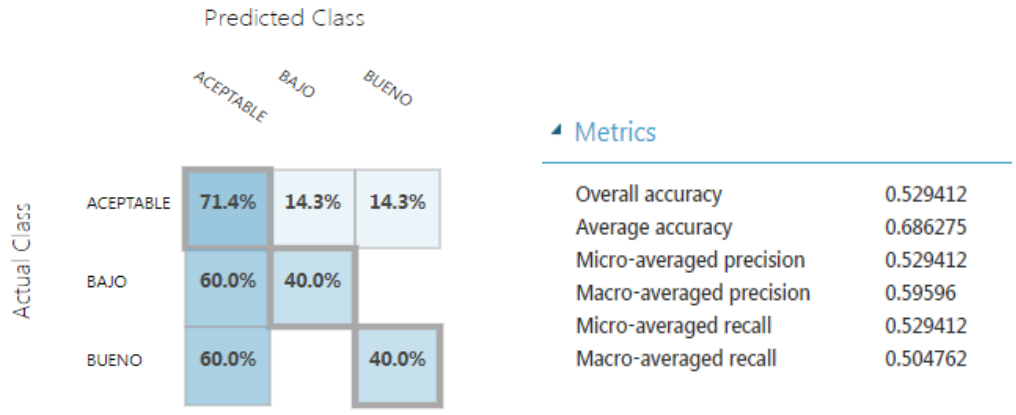


Ilustración 37. **Matriz de confusión y métricas del modelo empleando el método Hold-out y Bosque de Decisión Multilineal**

Fuente: Elaboración propia basada en los resultados de la plataforma de Microsoft ML Studio

Método K-Fold & regresión logística multilineal

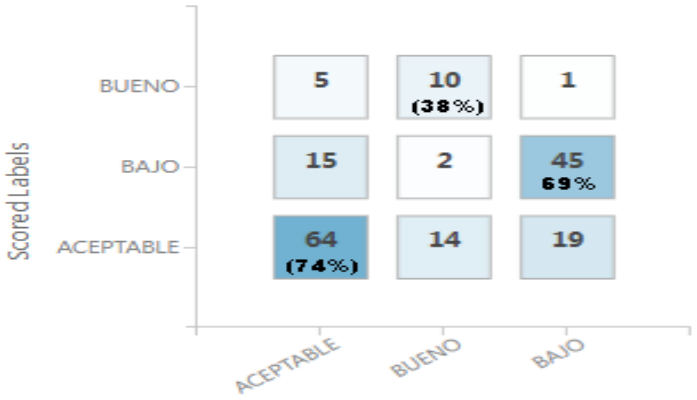


***Ilustración 38*. Matriz de confusión del modelo empleando el método K-Fold y la regresión logística multilineal**

Fuente: Elaboración propia basada en los resultados de la plataforma de Microsoft ML Studio

Método K-Fold & Bosque de Decisión Multilineal

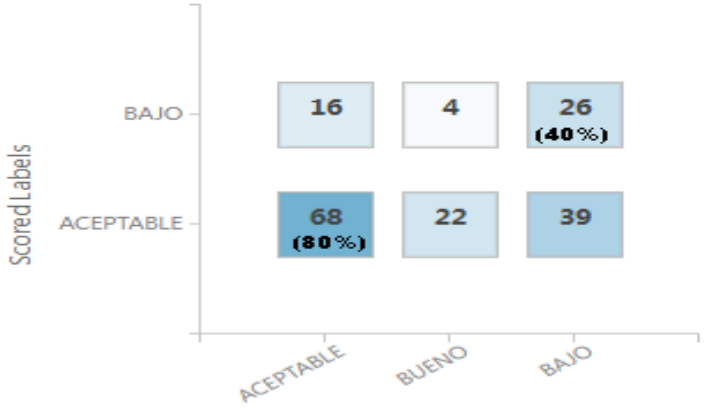


***Ilustración 39*. Matriz de confusión del modelo empleando el método K-Fold y Bosque de Decisión Multilineal**

Fuente: Elaboración propia basada en los resultados de la plataforma de Microsoft ML Studio

Como se observa el modelo que presenta una mayor precisión es el que emplea método Hold-Out y el método de Regresión Logística Multilineal. De acuerdo con los parámetros de la matriz de confusión y la estimación de un 70% para que sea un modelo usable (Maheshwari, 2015a). Por lo tanto, se convierte en el resultado que mejor se aproxima a los objetivos planteados del modelo propuesto.

4.5.2 Interfaz de la aplicación del modelo (API)

Microsoft Azure Machine Learning Studio ofrece la posibilidad de realizar una interfaz de programación de aplicaciones (API) con el modelo construido. Al emplear esta herramienta se obtiene la interfaz para la ejecución del modelo ingresando la respectiva serie de datos de entrada

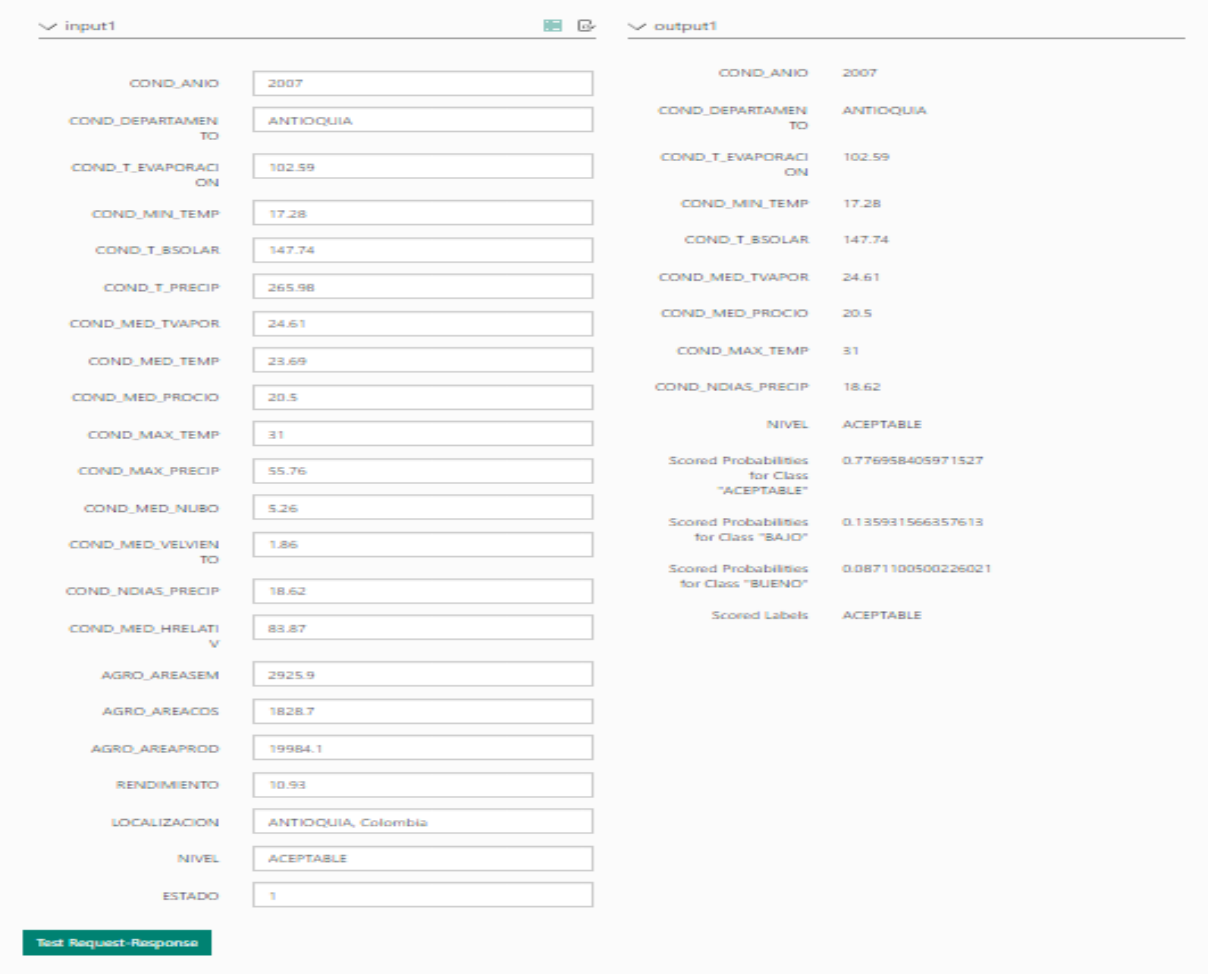


***Ilustración 40*. Interfaz para alimentar y probar el modelo construido**

Fuente: Elaboración propia

4.6 Implantación

4.6.1 Plan de implementación

Antes de proponer un plan de implementación del modelo, es preciso hacer en énfasis de las características de las organizaciones agrícolas:

Son organizaciones que en su mayoría presentan una deficiente infraestructura de Tecnologías de Información y Comunicaciones (TIC) y que como muchas organizaciones y empresas de otros sectores se encuentran en desventajas con empresas extranjeras en asuntos de inversión y apropiación tecnológica. Por lo tanto, sus esfuerzos deben enfocarse en las asociaciones o gremios por mantener informado tanto a sus miembros como demás productores de la región y socializar los riesgos y estrategias para el beneficio común del sector en materia de producción.

A continuación, se propone un plan de actividades para la aplicación del modelo en un terreno que se tenga la intención de cultivar Aguacate Hass.

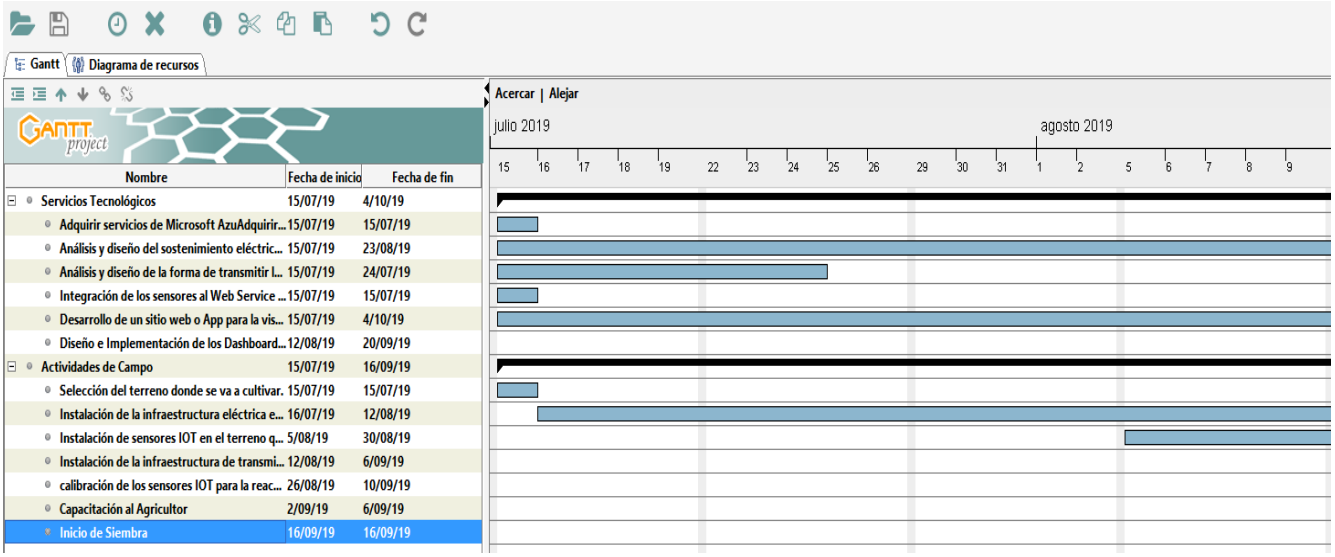


***Ilustración 41*. Plan de actividades para iniciar siembra**

Fuente: Elaboración propia

Tabla 19. Estructura de Costos por Hectárea

Costos por Hectarea		
Actividad	**Valor**	**%**
Mano de Obra	$ 3.750.000	17
Insumos, Equipos y Herramientas	$ 12.000.000	55
Dispositivos	$ 3.500.000	16
Microsoft Azure (Mes)	$ 600.000	3
Indirectos	$ 2.150.000	10
Costo Total	$ 22.000.000	100

4.6.2 Plan de monitoreo y mantención

Para el monitoreo y mantención del modelo se definen las siguientes estrategias:

- Revisión de las variables de impacto: cada año se debe revisar si las variables de impacto se mantienen corriendo el modelo Chi Cuadrado para confirmar si las variables de impacto se mantienen. Para confirmar este proceso es necesario diligenciar un documento de las siguientes características y quede evidencia del proceso.

Logo	Cultivo		
	Departamento	Formato	
	Fecha de Medición	Pag de	
Resultados			

Realizado Por

- Calibración del modelo periódicamente: Una vez realizada la revisión de las variables de impacto, se realiza la evaluación del modelo y si es el caso se debe ajustar la calibración del mismo con el objetivo de buscar la mejor precisión del modelo. Para llevar el seguimiento de la ejecución del proceso se debe diligenciar el siguiente formato.

Logo	Cultivo		
	Departamento	Formato	
	Fecha de Medición	Pag de	
Ajustes a realizar			
Realizado Por			

- Configuración de alertas: Dentro de la programación de los sensores se debe tener en cuenta envió de alertas de mal funcionamiento.
- Web Services de Monitoreo de Sensores: Se debe desarrollar un Web Services que este monitoreando los sensores activos del terreno en términos de comunicación y actividad, sí existe algún inconveniente se debe reportar vía alertas la novedad.
- Pareto con los resultados obtenidos: una vez se tengan datos de producción con el esquema implementado se debe realizar un informe Pareto con el rendimiento obtenido y el rendimiento del modelo.

CAPITULO 5. DISCUSIÓN DE RESULTADOS

- Existen muchas más variables relacionadas con el rendimiento en la producción agrícola del aguacate Hass. Sin embargo, se seleccionaron las variables meteorológicas utilizando la disponibilidad de los datos como medida clave para este proyecto de investigación.
- A partir de los datos del conjunto de variables meteorológicas comparadas y filtradas aplicando la prueba de chi-cuadrado, basado en la correlación de las características más significativas en relación al rendimiento se considera, según los datos, que las variables de mayor impacto son la temperatura en su valor mínimo y máximo, el total de evaporación, la media de tensión del vapor y el número de días de precipitación. Este hecho corrobora la relación encontrada entre la sincronización del calor y la humedad, junto acumulación de calor, para predecir los rendimientos del maíz (Crane-Droesch, 2018) y los parámetros óptimos de temperatura y precipitación para la mayor producción de trigo (Majumdar, Naraseeyappa, & Ankalaki, 2017). Sin embargo, aunque la hipótesis indica que las variables

meteorológicas de mayor impacto en el rendimiento son temperatura, humedad relativa y precipitación la hipótesis no puede ser descartada. Posiblemente, esta situación sucede por dos factores. En primer lugar, como consecuencia del estado de la información y los volúmenes limitados de datos periódicos que conllevan a realizar ajustes por medio de técnicas estadísticas para correlacionar los distintos parámetros. En segundo lugar, por la ubicación de cierto número de estaciones meteorológicas en zonas distantes de los ejes rurales de producción agrícola de aguacate.

- Por otro lado, de acuerdo con la matriz de confusión, la cual es una herramienta estadística usada para medir la precisión de los modelos predictivos; el modelo que presenta una mayor precisión es el que emplea método "Validación en Espera" y el método de "regresión logística multilineal", con valores mayores 70% de estimación. Esto implica que, con este modelo, es posible determinar con cierto grado de precisión, el estado del rendimiento de un cultivo a partir de las variables climáticas capturadas en un periodo o ciclo de producción siempre y cuando se mantenga el control de los demás factores que impactan en el rendimiento. Pero más allá de las características del modelo, cabe destacar dos aspectos principales: la metodología y la herramienta. Con la metodología llevada a cabo es posible construir nuevos modelos predictivos a partir de distintos conjuntos de datos que se ajusten a cada organización productora agrícola en particular de forma más precisa y robusta. Las herramientas para el procesamiento de datos y aplicación de técnicas de aprendizaje de máquina son cada vez más asequibles y accesibles, de forma que pueden adaptarse a los requerimientos de cada organización sin recurrir a conocimientos profundos en infraestructura tecnológica, tecnologías informáticas o de desarrollo de software. En el caso particular de Microsoft Azure ML Studio, las configuraciones y modificaciones en la estructura de ejecución de un modelo de análisis predictivo se realizan con conexiones de

bloques y ajustes de los parámetros internos con el propósito de pasar del concepto a la implementación.

- A partir de los resultados se conciben en general dos planteamientos. El primero consiste en reconocer la importancia de la disponibilidad e integridad de los datos para las organizaciones productoras agrícolas para generar modelos robustos y confiables que brinden solución a las necesidades del negocio. El segundo comprende en identificar la apropiación de la tecnología como un proceso estratégico para obtener una ventaja competitiva, reconociendo su potencial como habilitador en los procesos productivos y de negocio conforme se hace más asequible y accesible.

CAPITULO 6. CONCLUSIONES

Con el desarrollo de este proyecto de investigación se logró obtener un modelo predictivo optimo con la capacidad de identificar y clasificar las variables que impactan directamente el rendimiento en la producción de aguacate Hass con base en los valores de las variables meteorológicas de entrada. Las variables identificadas de mayor impacto son la temperatura en su valor mínimo y máximo, el total de evaporación, la media de tensión del vapor y el número de días de precipitación.

Con la ayuda de los modelos predictivos de la analítica de datos es posible la predicción de resultados adversos en un panorama real y simulado como una herramienta para diseñar medidas de contingencias para potenciales escenarios.

Para poder implantar el modelo en una organización productora agrícola es preciso delinear un plan de actividades de aplicación del modelo y un plan de monitoreo y mantención que defina el esquema de actividades dentro de un marco general. Sin embargo, no existe un único esquema óptimo que se adecúe a cada problema u organización. Pero siguiendo la metodología correctamente es posible construir soluciones más robustas alineada con las estrategias y la caracterización de las problemáticas a nivel de organización y sector.

Asimismo, se hace evidente la importancia de los datos para las organizaciones de producción agrícola porque constituyen un impulso hacia la integración de procesos y equipos tecnológicos. Por lo tanto, deben asegurar su disponibilidad e integridad para que puedan continuar escalando hacia soluciones más sofisticadas dentro del paradigma del *Big Data* que les permita tomar decisiones más inteligentes y ser más competitivas. De esta forma, deben hacerse de tecnologías asociadas en la generación, captura, procesamiento y almacenamiento de sus propios datos de forma dinámica.

Las herramientas para el análisis de datos comienzan a ser una limitante menos y se proyectan continuamente para facilitar la toma de decisiones en distintas áreas del conocimiento. Actualmente, existen software especializado en análisis de datos, de código abierto y comerciales, para programadores o con mayor facilidad de uso, la elección depende de la necesidad de la investigación. Microsoft Machine Learning Studio es una herramienta con un entorno grafico que permite pasar rápidamente del concepto a la implementación. Por lo cual, se convierte en una herramienta apropiada para aplicaciones y usuarios empresariales que buscan simplificar la construcción de modelos de aprendizaje de máquina y brindar valor a los proyectos de análisis de datos en un tiempo corto. Por supuesto, sin prescindir del conocimiento y las habilidades en ciencia de datos.

De esta manera, se concuerda en la relevancia de gestionar talento con habilidades en ciencia de datos para incrementar las probabilidades de éxito en los proyectos de análisis de datos 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 mercados actuales.

REFERENCIAS

Agrawal, Y. (2019). Hypothesis testing in Machine learning using Python. Retrieved March 7, 2019, from https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

Agricultura Moderna (2018). El GPS, un aliado para la precisión. Obtenido de http://agmoderna.com.ar/tecnologia-en-el-campo/el-gps-un-aliado-para-la-precision/

Agronet, E. A. del M. de A. y D. R. (2007). Estadísticas Inicio - Agronet. 2017.

Aguilar, L. J. (2013). Big Data Análisis de Grandes Volúmenes de Datos en Organizaciones. Alfaomega Grupo Editor, ed.Primera. Ciudad de México: Alfaomega.

AINIA (2015). Drones, agricultura de precisión e industria alimentaria: Nuevas tendencias. Obtenido de http://www.ainia.es/tecnoalimentalia/tecnologia/drones-agricultura-de-precision-e-industria-alimentaria-nuevas-tendencias/

Ruiz, A., Basualdo, M. S, y Matich, J. D. (2001). Cátedra: Informática Aplicada a la Ingeniería de Procesos-Orientación I Redes Neuronales: Conceptos Básicos y Aplicaciones. Universidad del Rosario.

Ariza López, F. J., Rodríguez Avi, J., & Alba Fernandez, V. (2018). Control Estricto De Matrices De Confusión Por Medio De Distribuciones Multinomiales. Revista Internacional de Ciencia y Tecnología de La Información Geográfica, 215–227. Recuperado de https://doi.org/10.21138

Baoss Analytics Everywhere (2015). Las 4 V´s del Big Data - BAOSS.

Bigeek. (2018). DataWarehouse y DataMart: qué son y para qué sirve.

Bogue, R. (2017). Sensors key to advances in precision agriculture. Sensor Review, 37(1), 1–6. Recuperado de https://doi.org/10.1108/SR-10-2016-0215

Borzenkova, I. (2009). TYPES AND CHARACTERISTICS OF PRECIPITATION. Encyclopedia of Life Support Systems (EOLSS), II, 364.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? Ssrn. Recuperado de https://doi.org/10.2139/ssrn.1819486

Carillo, K. D. A., Galy, N., Guthrie, C., & Vanhems, A. (2018). How to turn managers into data-driven decision makers. Business Process Management Journal. Recuperado de https://doi.org/10.1108/BPMJ-11-2017-0331

CleverData (2018). ¿Qué es Machine Learning?

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(1), 270. Recuperado de https://doi.org/10.1186/s12859-018-2264-5

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003. Recuperado de http://stacks.iop.org/1748-9326/13/i=11/a=114003

Díaz, J. C. (2016). Organizaciones orientadas al dato / Transformando organizaciones hacia una cultura analítica. ed. 1ª. Barcelona: 2016.

Dinero (2017). Aguacate exportación y mercado en Colombia.

Faostat (2017). Estadísticas agrícolas de aguacate: producción, superficie y rendimiento.

Fariña, J. (1998). La ciudad y el medio natural.

Gallardo, J. A. (2009). Metodología para la definición de requisitos en proyectos de data mining.

GPS (2018). GPS.gov. Retrieved May 11, 2018. Recuperado de https://www.gps.gov/spanish.php

Griffin, P. A., & Wright, A. M. (2015). Commentaries on Big Data's Importance for Accounting and Auditing. Accounting Horizons, 29(2), 377–379. Recuperado de https://doi.org/10.2308/acch-51066

IBM (2011). IBM SPSS Modeler CRISP-DM Guide. IBM Corp.

IEBS (2016). ¿Cuáles son las 5 V's del Big Data? - Blog de IEBSchool.

Instituto de Hidrología Meteorología y Estudios Ambientales IDEAM. (2005). Parte I Aspectos Nacionales. In Atlas Climatológico de Colombia (1st ed., p. 18). Recuperado de http://documentacion.ideam.gov.co/openbiblio/bvirtual/019711/019711.htm

Jiawei Han, Micheline Kamber, J. P. (2011). *Data Mining – Concepts & Techniques*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

José, A., y Gallardo Arancibia, A. (2009). Departamento de Lenguajes y Sistemas informáticos Ingeniería de Software Facultad de Informática. Metodología para la Definición de Requisitos en Proyectos de Data Mining.

Kerley, C. (2017). What Are the Different Types of Correlations? | Sciencing. Retrieved August 14, 2018. Recuperado de https://sciencing.com/different-types-correlations-6979655.html

Kirasich, K., Smith, T. y Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review, 1(3), 25. Recuperado de https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Leal, E. J. H., Méndez, N. D. D., y Cadavid, J. M. (2017). Big Data: an exploration of research, technologies and application cases. TecnoLógicas, 20(39).

Maheshwari, A. K. (2015a). Business Intelligence and Data Mining. (M. Ferguson, Ed.).

Chennai: Business Expert Press.

Maheshwari, A. K. (2015b). Business Intelligent & Data Mining Made Accesible. Fairfield: ISTED, WILEY.

Majumdar, J., Naraseeyappa, S., y Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data, 4(1), 20. https://doi.org/10.1186/s40537-017-0077-4

Marr, B. (2016). Big data in practice : how 45 successful companies used big data analytics to deliver extraordinary.

McHugh, M. L. (2013). The chi-square test of independence. Biochemia Medica, 23(2), 143–149. Recuperado de https://doi.org/10.11613/BM.2013.018

Melker, A. I., Starovoitov, S. A., & Vorobyeva, T. V. (2010). Heat, Temperature, Entropy. Materials Physics and Mechanics , 9. Recuperado de http://www.ipme.ru/e-journals/MPM/no_3910/melker4.pdf

Microsoft. (2018). Algoritmo de árboles de decisión de Microsoft | Microsoft Docs.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). Introduction to linear regression analysis. Wiley. Recuperado de https://books.google.com.co/books?id=27kOCgAAQBAJ&dq=linear+regression+analysis&hl=es&source=gbs_navlinks_s

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal : The Journal of Medical Association of Malawi, 24(3), 69–71. Recuperado de http://www.ncbi.nlm.nih.gov/pubmed/23638278

Norte de Santander, con grandes posibilidades para proveer aguacate. (2018, December 10). La Opinión. Recuperado de https://www.laopinion.com.co/economia/norte-de-santander-

con-grandes-posibilidades-para-proveer-aguacate-156745#OP

Philipp, G., & Carbonell, J. G. (2017). Nonparametric Neural Networks (pp. 1–28). Retrieved from http://arxiv.org/abs/1712.05440

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Recuperado de https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Posada, A. (2017). El Internet de las Cosas y la Agricultura de Precisión. Recuperado de http://techcetera.co/internet-las-cosas-la-agricultura-precision/

ProColombia. (2017). Acceso y oportunidades para la exportación del aguacate Hass Colombiano al mercado de los Estados Unidos.

Rodriguez, J. R. (2016). ¿Cómo son las empresas orientadas a los datos? *Harvard Deusto Business Review*, 256 (27), pp. 46–54.

Rouse, M. (2012). ¿Qué es Análisis de datos? Retrieved September 6, 2018. Recuperado de https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos

Russom, P. (2011). Big data analytics. TDWI BEST PRACTICES REPORT. Recuperado de https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

Salcedo Parra, O. J., Galeano, R. M., & Rodriguez, L. G. (2010). Metodología crisp para la implementación Data Warehouse.

Sawyer, S. F. (2013). Analysis of Variance: The Fundamental Concepts. The Journal of Manual & ManipulaTive Therapy N, 17(2), 27E-38E. Recuperado de http://jmmtonline.com/documents/v17n2/sawyer.pdf

Serrano, E. (2017). Big Data y sector Agro: un binomio de futuro. Recuperado de http://www.agrointeligencia.com/big-data-agricultura-enrique-serrano/

Sinnexus (2018a). ¿Qué es Business Intelligence?

Sinnexus (2018b). Datawarehouse.

Sonka, S. (2016). Big Data: Fueling the Next Evolution of Agricultural Innovation. Journal of Innovation Management, 4(1), 114–136. Recuperado de https://journals.fe.up.pt/index.php/IJMAI/article/view/163#.XDVoKL05JxQ.mendeley

Strange, R., & Zucchella, A. (2017). Industry 4.0, global value chains and international business. Multinational Business Review, 25(3), 174–184. Recuperado de https://doi.org/10.1108/MBR-05-2017-0028

Swalin, A. (2018). Choosing the Right Metric for Evaluating Machine Learning Models — Part 2. Recuperado de https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428

Timm, N. H. (2007). Applied multivariate analysis. Springer. Recuperado de https://books.google.com.co/books?id=vtiyg6fnnskC&dq=multilinear+regression+analysis&hl=es&source=gbs_navlinks_s

Ullrich, C. (2009). Forecasting and hedging in the foreign exchange markets. Springer-Verlag. Recuperado de https://books.google.com.co/books?id=zs1CpfZeDgcC&dq=cross+validation&hl=es&source=gbs_navlinks_s

Valencia, E. (2015). Data Coaching Un librito sobre datos y beneficios. Leanpub, ed. Spanish.

Ylijoki, O., & Porras, J. (2018). A recipe for big data value creation. Business Process Management Journal, BPMJ-03-2018-0082. https://doi.org/10.1108/BPMJ-03-2018-0082

Zheng, A. (2015). Evaluating Machine Learning Models - O'Reilly Media. Recuperado de https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/4/offline-

evaluation-mechanisms-hold-out-validation-cross-validation-and-bootstrapping