UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

TEMA:

Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.

AUTOR:

Zambrano Yont Cristhian Oswaldo

Trabajo de titulación previo a la obtención del título de

INGENIERO EN SITEMAS COMPUTACIONALES

TUTOR:

Ing. Castro Aguilar Gilberto Fernando

Guayaquil, Ecuador

13 de marzo del 2021

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENNIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación, fue realizado en su totalidad por **Zambrano Yont Cristhian Oswaldo**, como requerimiento para la obtención del título de **Ingeniero en Sistemas Computacionales**.

TUTOR

f. _____
Ing. Castro Aguilar Gilberto Fernando

Guayaquil, a los 13 días del mes de marzo del año 2021

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENNIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, Zambrano Yont Cristhian Oswaldo

DECLARO QUE:

El Trabajo de Titulación, **Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.** previo a la obtención del título de **Ingeniero en Sistemas Computacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 13 días del mes de marzo del año 2021

EL AUTOR

f. _____

Zambrano Yont Cristhian Oswaldo

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENNIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

AUTORIZACIÓN

Yo, Zambrano Yont Cristhian Oswaldo

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 13 días del mes de marzo del año 2021

EL AUTOR:

f. _____

Zambrano Yont Cristhian Oswaldo

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD FACULTAD DE INGENIERIA
CARRERA INGENIERIA EN SISTEMAS COMPUTACIONALES
PERIODO II-2020 UTE

ACTA DE TRIBUNAL DE SUSTENTACIÓN
TRABAJO DE TITULACIÓN

En sesión del día 13 de Marzo de 2021, el Tribunal de Sustentación ha escuchado y evaluado el Trabajo de Titulación denominado "DISEÑO DE UN MODELO PREDICTIVO, MEDIANTE LA TÉCNICA DE MINERÍA DE DATOS PARA LA ASIGNACIÓN DE RECURSOS EN LA PRODUCCIÓN DE CAFÉ SOLIDO SOLUBLE PARA CALIDAD A/R DE LA COMPAÑÍA ASKELGADO S.A.", elaborado por el/la estudiante CRISTHIAN OSWALDO ZAMBRANO YONT, obteniendo el siguiente resultado:

Nombre del Docente-tutor	Nombres de los miembros del Tribunal de sustentación		
GILBERTO FERNANDO CASTRO AGUILAR **Etapas de ejecución del proceso e Informe final** **Nota sobre 10:** 9.0	ANA ISABEL CAMACHO CORONEL Nota sobre 10: 8.0 Total: 20 %	JORGE SALVADOR PESANTES MENDEZ Nota sobre 10: 8.0 Total: 50 %	GALO ENRIQUE CORNEJO GOMEZ Nota sobre 10: 7.5 Total: 30 %
Parcial: 50 %	Parcial: 50 %		
Nota final ponderada del trabajo de titulo:			

Para constancia de lo cual los abajo firmantes certificamos.

Miembro 1 del Tribunal Miembro 2 del Tribunal Oponente

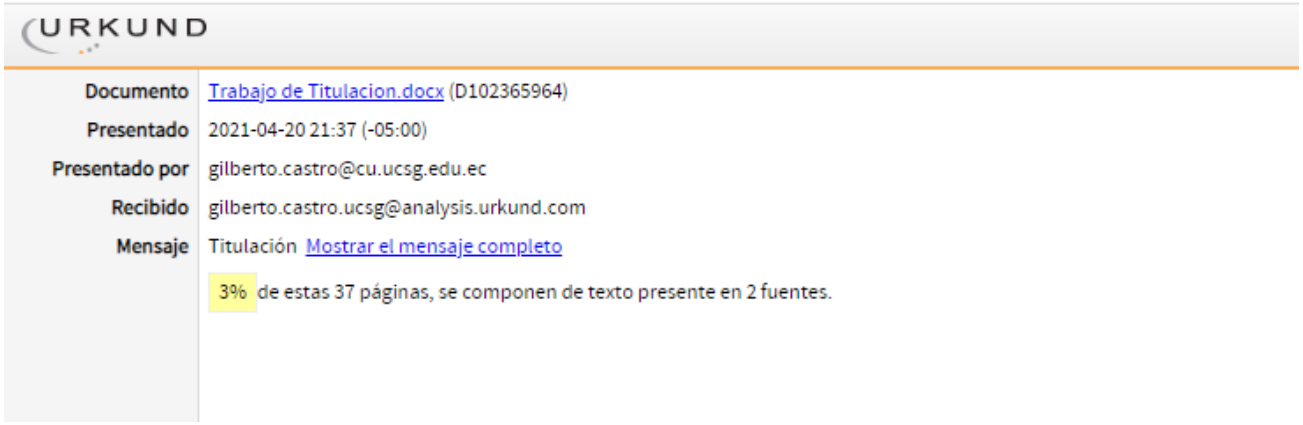
Docente Tutor

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENNIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

REPORTE URKUND



URKUND

Documento	Trabajo de Titulacion.docx (D102365964)
Presentado	2021-04-20 21:37 (-05:00)
Presentado por	gilberto.castro@cu.ucsg.edu.ec
Recibido	gilberto.castro.ucsg@analysis.urkund.com
Mensaje	Titulación Mostrar el mensaje completo
	3% de estas 37 páginas, se componen de texto presente en 2 fuentes.

TUTOR

f. _____

Ing. Castro Aguilar Gilberto Fernando

AGRADECIMIENTO

Mis más sinceros agradecimientos a mi familia, amigos y al Ing. Fernando Castro por el apoyo profesional y la paciencia para guiarme a lo largo de mi trabajo de titulación y a todos los docentes que han formado parte de este proceso de formación educativa.

DEDICATORIA

Este trabajo de titulación es dedicado a mi familia y amigos, los cuales han sido un motor fundamental en el desarrollo y finalización de mi carrera universitaria.

ÍNDICE

RESUMEN

El presente trabajo de titulación del desarrollo de un diseño de un modelo predictivo para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A. tiene como objetivo identificar los recursos que se van a usar, reconocer una técnica de minería de datos, usar un modelo predictivo y evaluar el modelo para la calidad A/R. La investigación utilizada fue del tipo documental, ya que se realizó una entrevista a los 2 responsables de la producción para conocer su estado actual frente al tema de asignación de recursos. Para el proyecto se propuso el diseño de un modelo predictivo, mediante la técnica de minería de datos para identificar los recursos mediante históricos. Para realizar la investigación se utilizó un enfoque de investigación cualitativo, de tipo analítico para estudiar el contexto en donde existe el problema de la identificación de recursos; La técnica de árboles de decisión son técnicas de aprendizaje supervisado, en las que se aprenden funciones, relaciones que asocian entradas con salidas, por lo que se ajustan a un conjunto de ejemplos de los que conocemos la relación entre la entrada y la salida deseada. La técnica de minería es supervisada, para eso se usará python y la metodología es CRISP-MD. Se diseñó el un árbol que genero un modelo predictivo de evaluación para la toma de decisiones en la planta. Al final se plantearon las recomendaciones a considerarse como la mejora del modelo previamente utilizado.

Palabras Clave: Minería de datos, Python, Árbol de decisión, Modelo Predictivo.

ABSTRACT

The present qualification work of the development of a design of a predictive model for the allocation of resources in the production of soluble solid coffee for A / R quality of the company ASKELGADO S.A. aims to identify the resources to be used, recognize a data mining technique, use a predictive model, and evaluate the model for A / R quality. The research used was of the documentary type, since an interview was conducted with the 2 people in charge of the production to find out their current status regarding the allocation of resources. For the project, the design of a predictive model was proposed, using the data mining technique to identify resources through historical data. To carry out the research, a qualitative, analytical research approach was used to study the context where the problem of resource identification exists; The decision trees technique are supervised learning techniques, in which functions are learned, relationships that associate inputs with outputs, so they conform to a set of examples of which we know the relationship between the input and the desired output. The mining technique is supervised, for that python will be used and the methodology is CRISP-MD. A tree was designed that generated a predictive evaluation model for decision making in the plant. In the end, the recommendations were made to be considered as the improvement of the previously used model.

Key words: Data Mining, Python, Decision Tree, Predictive Model.

INTRODUCCIÓN

Cada vez es más difícil ignorar que el mundo se encuentra rodeado de información, la misma que puede ser usada para la recolección o extracción de los datos dentro de muchas áreas como el conocimiento, estadística, finanzas, negocios, medicina, agricultura e industria, es tanto así, que la, humanidad ya había generado 281 exabytes en el año 2007, lo dijo Eric Schmidt, CEO de Google, en una conferencia en el 2010 y se estimaba que 4 años más adelante se llegarían a los 1800 exabytes de información a nivel mundial con un crecimiento exponencial (*¿Cuánta Información y Datos Generamos al Año En El Mundo?*, 2013.). Todas estas variables hacen que la tecnología tenga que avanzar al mismo paso que se va generando la data, es por eso que cada vez más los algoritmos para la búsqueda de información como los programas informáticos que se usan para entregar respuestas exactas (Aveiro, 2019.) se vuelven complejos y necesarios en un mundo que crece constantemente.

En la última década se han empezado a implementar técnicas de búsqueda de información que nos permita buscar respuestas de manera eficiente y eficaz, debido a que hay una mayor demanda de productos y servicios dentro de los mercados, haciendo que los usuarios quieran soluciones inmediatas, esto ha hecho que los dueños o gerentes de negocios tengan que tomar decisiones más rápidas para poder generar resultados en menos tiempo.

Debido a estas necesidades se han creado técnicas que permiten crear cuadros de tendencias a través de estadística y de históricos, como lo es la minería de datos, que permite que la recolección de datos de manera tradicional pase a segundo plano, ahorrando mucho tiempo a través de la creación de modelos predictivos o descriptivos a partir de análisis de muestras de datos o registros.

La Minería de Datos mezcla e integra técnicas de diferentes disciplinas como tecnologías de bases de datos y data warehouse, estadística, aprendizaje de máquina, computación de alta performance, computación evolutiva, reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, y

análisis de datos espaciales o temporales. Como una sub área específica de la Minería de Datos se puede decir que el Data Stream Mining que es el proceso de extraer conocimiento en estructuras de datos continua. (Schab et al., 2018). La toma de estas decisiones va acompañada de simulaciones que se pueden hacer bajo los modelos predictivos, que son un sistema que emplea datos y estadística para predecir resultados a partir de unos modelos de datos; desde resultados deportivos y audiencias televisivas hasta avances tecnológicos y ganancias empresariales.

El modelado predictivo se suele conocer también como:

- Análisis predictivos o Analítica predictiva.
- Aprendizaje automático.

(*Modelado predictivo*, n.d.)

Análisis predictivo

El análisis predictivo usa datos históricos para predecir eventos. Estos se utilizan para crear un modelo matemático que capturen las tendencias importantes. Este modelo predictivo se usa con los datos actuales para predecir lo que pasará, o bien para incitar acciones que se llevan a cabo con el fin de obtener resultados óptimos.

Muy constantemente se habla del análisis predictivo en el contexto del big data, cada vez más, los negocios toman decisiones basadas en los datos procedentes de esta valiosa mina de información. (*Análisis predictivo*, 2020.)

A pesar de las diferencias metodológicas y matemáticas entre los tipos de modelos, el objetivo general de todos ellos es similar: **predecir resultados futuros** basándose en datos pasados. (Spain, 2020a)

Aprendizaje automático

Es el subconjunto de inteligencia artificial (IA) que se centra en desarrollar sistemas que aprenden, o mejoran el rendimiento, en función de los datos que consumen. Inteligencia artificial es un término amplio que se refiere a sistemas o máquinas que imitan la inteligencia humana. Se suele mencionar al aprendizaje autónomo y a la IA en las mismas conversaciones,

pero no significan lo mismo. Un aspecto importante a destacar es que, aunque todo aprendizaje autónomo es IA, no toda IA es aprendizaje autónomo, esto funciona en todo a nuestro alrededor. (*¿Qué Es El Aprendizaje Automático?* | *Oracle Chile*, 2014.)

El café es uno de los productos más consumidos a nivel mundial, es el segundo más comprado después del agua, ha tenido una gran relevancia y a grandes rasgos el café se divide en 2 sectores: de consumo general y especiales.

El café de consumo general representa al 90% del café que se comercializa en todo el mundo, desde el punto de vista del volumen. El desequilibrio entre la oferta y la demanda penaliza a los cafés de consumo general. La calidad de un lote de café de este sector depende principalmente de criterios materiales:

- Una tasa de humedad de entre 11% y 12%.
- El tamaño o la calidad del grano, con una tolerancia de un pequeño porcentaje de granos de calidad inferior en un lote de una calidad determinada.
- En cada tasa se atribuye un valor de defectos determinado.

Una tasa aceptable de defectos en cada calidad. A cada defecto se atribuye un valor determinado (granos negros, granos dañados, granos rotos, cuerpos extraños, etc.). Por ejemplo, la Green Coffee Association, de Nueva York, cuenta 10 granos dañados por la broca del café como un defecto, mientras que un sólo grano negro cuenta como un defecto. (*Good Hygiene Practices*, 2015.)

La calidad es un término que se utiliza en todo el mundo, es la capacidad que posee un objeto para poder satisfacer necesidades implícitas o explicitas según se requiera (*Significado de Calidad*, 2018)

Cuando se va a producir un café siempre se le realiza un análisis físico como son la separación de impurezas que comprenden las piedras, cascara, cereza y granos que se encuentren perforados o quebrados, una vez identificada sus impurezas se hace un cálculo de diferencia de pesos en porcentajes, de los cuales se toma una cierta cantidad de granos que es

pesada y se separa la diferencia con las impurezas. Varios de los parámetros que determinan la calidad son:

- El olor que debe ser característico a café seco y fresco.
- El porcentaje de humedad, que debe estar entre el 10.5% al 12%.
- Merma o pergamino que son partes de cascara seca que quedan.
- Tamaño del grano.
- Los defectos que son las modificaciones que ha tenido el grano, estos pueden ser café brocado o quebrados, de los cuales se permite del 0.5% al 1.5%.

  (Posada, 2019).

En la compañía existen varios tipos de calidades de café soluble de las cuales se encuentran: 77/S2, 77/S1, A2, R2, A/R, cabe recalcar que estas calidades son específicas de la compañía ya que cada una de estas se encuentran diseñadas de acuerdo a los requerimientos de cada cliente. Con esta base de evaluación, un café arábico sin lavar de calidad 2 del Brasil no debe contener más de 6 defectos como granos negros, cereza, piedras, palos, pergamino como la capa más fina de la cascara, cascara, café brocado, en otras palabras, con fisuras por polillas, mientras que un café de calidad 5 del mismo origen puede tener hasta 60 defectos en una muestra de 16 pulgadas cúbicas de volumen.

El café Especial representa el 10% del volumen del suministro mundial del café. Se trata de un mercado muy limitado y la venta de este producto sigue siendo limitada a pesar de haber crecido. En pocas palabras, estos cafés proceden de plantaciones claramente determinadas que trabajan por contrato con la industria del café, o de organizaciones de productores que tienen contratos con la industria o venden su café en subastas.

Los cafés especiales tienen que satisfacer criterios de selección muy estrictos, así como criterios de calidad organoléptica excepcional. Por ejemplo, los cafés kenyanos de mayor calidad destacan por la calidad de la bebida que producen.

El precio de estos cafés no tiene nada que ver con la Bolsa de New York. Solo unos pocos cafés de origen especial se venden bajo contrato, como el café Blue Mountain de Jamaica, el café de las Islas Galápagos o algún café Arábica de Aceh, Indonesia, y el precio de café de mayor calidad toma en cuenta la bolsa. El precio viene de Nueva York.(*Good Hygiene Practices*, n.d.).

La calidad A/R es una calidad diseñada por la compañía ASKELGADO S.A. específicamente para clientes de diferentes partes de Europa como Inglaterra, España, Alemania, Holanda, Italia y Republica Checa, esta calidad consiste en la combinación del café arábica y robusta que están divididos en 50% café Arábica y 50% café Robusta, los cuales están regidos bajo múltiples cambios en parámetros dependiendo del cliente, estos cambios pueden variar entre la densidad, color y ph del café.

Este tema es importante ya que las industrias siempre están en mejoras constantes, sobre todo las del sector alimenticio, se encuentran muy involucradas en el estricto cuidado de sus procesos, como se mencionó previamente en el documento, con la llegada del Covid-19 (la pandemia) el consumo del café cada vez se hace más presente, ya que el consumo casero de este producto se ha aumentado y las empresas de la industria del café aumentaron sus ventas, por eso la importancia de un modelo predictivo que permita decidir de manera más rápida y efectiva.

El presente documento consiste en la creación y testeo de un modelo predictivo, cuya estructura del documento contiene:

- Capítulo I: El planteamiento del problema a tratar, donde se definen la ubicación del problema en conjunto con la formulación y evaluación, así mismo se plantearán los objetivos, generales y específicos, el alcance, justificar e importancia, hipótesis y las variables de investigación.
- Capítulo II: Marco teórico, en donde se recopilan los antecedentes, investigaciones previas y todas las consideraciones teóricas del trabajo de investigación.
- Capítulo III: Metodología de la investigación, también se va a definir la población, muestra, tamaño de la muestra, los instrumentos de recolección de datos, los instrumentos de investigación y las encuestas o cuestionario.

- Capítulo IV: Propuesta tecnológica como las herramientas, el análisis de datos, el criterio de toma de decisiones, las técnicas para procesamiento y análisis de datos y modelo entidad relación.
- Conclusiones: donde se concluye todo el trabajo de investigación.

Es por ello que el presente trabajo de titulación busca proponer el Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.

CAPÍTULO I

EL PROBLEMA

PLANTEAMIENTO DEL PROBLEMA

Ubicación del Problema en un Contexto

En la empresa de café ASKELGADO S.A. la toma se de decisiones para la asignación de recursos para la fabricación del café solido soluble se lo hace a través de un método llamado ruta que consiste en el análisis de los requerimientos del cliente y el análisis de los históricos para poder asignar de manera manual los recursos y parámetros que este debe cumplir, a través de fórmulas pre planteadas se realiza la asignación de la materia a usar, para esto el enfoque principal del problema se basa específicamente en la calidad A/R, la cual tiene algunas variaciones, si bien es cierto es la misma pero varios cambios. En el café existen diferentes tipos de grano como Arábica, Robusta, Kopi Luwak (*La cosecha y los tipos de granos de café*, n.d.). La particularidad de A/R es que solo cuenta con 2 tipos de granos que son Arábica y Robusta los cuales son combinados en porcentajes iguales.

El proceso debe ser consistente con las metas y requerimientos internos y externos, lo que ayuda a establecer formas efectivas de obtener más liquidez y reducir los gastos, porque la liquidez es apoyo financiero. De la empresa. (administrador, 2019).

Es por esto que las empresas industriales, sobre todo las de alimentos o commodities como el café se encuentran en una carrera contra el tiempo para poder mejorar constantemente sus procesos y optimizar sus recursos en el menor tiempo posible, ya que existe una demora en la toma de decisiones y asignación de recursos que podría ser mejorada.

Causas y Consecuencias del Problema

De acuerdo con lo indicado por el jefe de control de calidad de la empresa ASKELGADO S.A., los métodos aplicados para el cálculo y las asignaciones de los recursos toman un tiempo de una media hora en lo que se buscan los requerimientos, se realiza el ingreso de los mismos datos y se asigna la ruta o cálculo de recursos realizados por medio de una hoja de

cálculo, para conocer esto anticipadamente existen históricos, que en su gran mayoría se encuentran en papel y eso es una de las principales causas por la que se ha llevado de esta manera, además siempre ha existido el experto encargado de esta área que maneja sus datos parcialmente de manera tecnológica, por lo que la empresa requiere mejorar eso, esto también ha generado una dependencia absoluta de dicha persona, siendo un obstáculo para la optimización de los tiempos en la asignación de recursos, ya que si esta persona no se encuentra, el proceso puede demorar más de lo esperado.

Delimitación del Problema

El modelo predictivo propuesto para el trabajo de titulación será generado con datos de estudio de los últimos 2 años de producción, los cuales están enfocados a la producción de café soluble de la calidad A/R.

Campo:	Tratamiento de datos en la producción de café
Área:	Minería de datos
Aspecto:	Un modelo predictivo para medir los recursos en la producción de café
Tema:	Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A**.**

Formulación del Problema

En la actualidad hay una alta demanda de café a nivel mundial y más en estos últimos años, lo cual está obligando a los países productores de café, los cuales antes procesaban el café como materia prima a construir nuevas fábricas haciendo que el exista una competencia muy fuerte dentro de los mercados y logrando que el café en especial el sólido soluble deba salir más rápido hacia los mercados, produciendo una búsqueda de nuevas formas para optimizar procesos y reducir costes para poder optimizar el tiempo y los beneficios.

Por lo cual para la formulación del problema debe responder las siguientes preguntas, ¿Cómo la técnica de minería de datos puede apoyar a la mejora de los procesos? ¿Cómo se puede mejorar la toma de decisiones para la asignación de recursos? ¿Qué herramientas son necesarias para poder ejecutar el modelo planteado?

Evaluación del Problema

Para la evaluación del problema se tomarán en cuenta las siguientes características:

Delimitado: El problema está delimitado a la identificación de variables que permitan realizar el análisis de los datos para la proyección de la información necesaria para poder identificar los recursos que se van a usar en la producción del café soluble, específicamente en la calidad A/R.

Claro: Se va a desarrollar mediante una técnica de minería de datos que va a permitir pronosticar los recursos a usar en la producción de café soluble, las técnicas se pueden ajustar al modelo predictivo para su desarrollo, tales como el árbol de decisión y regresión, curva adaptativa, análisis de duración.

Concreto: Es concreto porque se va a usar la técnica de minería de datos para la producción de café soluble en la calidad A/R para la empresa ASKELAGADO S.A.

Relevante: Es relevante ya que mediante la técnica de minería de datos se puede innovar en el campo del sector industrial cafetalero.

Factible: Es factible el diseño del modelo predictivo porque las áreas para la infestación son muy específicas en el campo industrial de la producción de café solido soluble y en recursos se cuenta con el acceso a la información

necesaria sobre los recursos que se van a usar para la producción de la calidad A/R.

Identifica los productos esperados: Es útil ya que la creación de este modelo predictivo abre un nuevo campo para la investigación de la implementación de la minería de datos en el sector cafetalero.

OBJETIVOS

OBJETIVO GENERAL

Diseñar un modelo predictivo aplicando una técnica de minería de datos para la empresa ASKELGADO S.A. en la asignación de recursos en la producción de café sólido soluble de la calidad A/R.

OBJETIVOS ESPECÍFICOS

- Identificar los recursos que se van a usar para la producción de la calidad A/R.
- Reconocer una técnica de minería de datos que se pueda ajustar al modelo predictivo.
- Usar un modelo predictivo aplicando una técnica de minería de datos que aproveche los recursos identificados en la investigación.
- Diseñar el modelo para la asignación de recursos de la producción de café solido soluble para calidad A/R es válido.

ALCANCES DEL PROBLEMA

El modelo predictivo por desarrollar será logrado a través de la técnica CRISP-DM y herramientas que permitan realizar la minería de datos que se ajusten de tal manera que permita la asignación de recursos para la producción de café solido soluble para la calidad A/R. Para esto también se debe lograr:

- Recolección de la información archivada en históricos.
- Digitalización de los datos con información organizada.

- Identificación de la técnica de minería de datos.
- Uso de herramientas como jupyter para programación en Python.
- Aplicación de las técnicas de minería para el modelo predictivo.
- Generar datos para la toma de decisiones.

JUSTIFICACION E IMPORTANCIA

El propósito de esta investigación es solucionar los problemas mencionados anteriormente mediante una herramienta de análisis predictivo para la asignación de recursos para la compañía ASKELGADO S.A., usando la técnica de minería de datos para el análisis de los históricos que existen sobre la producción del café solido soluble en la calidad A/R. Los resultados podrán ser utilizados para la optimización en la toma de decisiones y asignación de recursos.

Esto es principalmente para apoya al modelo predictivo. Dado un universo de estudio, el objetivo de un sistema de reconocimiento de la data que consiste en particionar dicho universo en clases, de tal manera que el sistema asignará a un elemento “x” el cual contendrá las variables predictoras y un elemento Y que va a contrastar. Es decir, el problema es de regresión, el sistema reconoce que un elemento pertenece a una clase.

Partes del problema:

- Aislar los objetos.
- Extraer los valores de las variables.
- Construir el sistema de regresión definiendo una distancia que permita medir predecir cuáles serán los valores contenido dentro de cada variable.

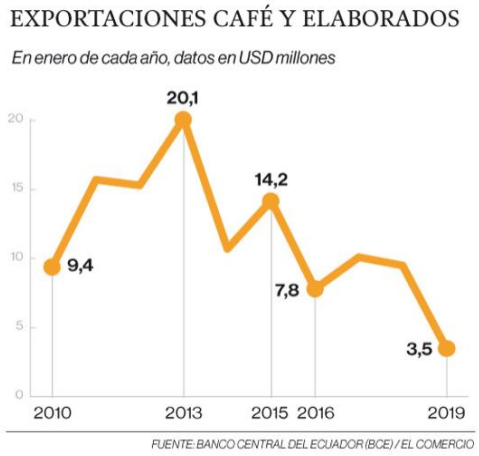
La ventaja de este modelo es que se ha dado en otras industrias y en varios países en los cuales el conocimiento y la experiencia con la Minería de datos es mayor a la que hay en Ecuador desde un contexto más general, por lo que para las empresas industriales dentro del Ecuador esto puede despertar un mayor interés e importancia en la implementación de recursos tecnológicos como lo es minar los datos, es muy necesario en estos tiempos

por la alta competitividad que existe en los mercados y la gran cantidad de información que se genera diariamente.

La industria del café ha sido y es muy importante ya que el café forma parte de la vida cotidiana de la mayoría de personas a nivel mundial, Ecuador es uno de los países que produce y fabrica café soluble a nivel mundial, el cual es distribuido a muchas partes del mundo, compitiendo con países como Brasil, Vietnam, Colombia, Indonesia, Honduras, Etiopia, India, Uganda, México y Perú, generándole al Ecuador ganancias mayores a los 66 millones de dólares anualmente de ingresos por exportaciones anualmente, estos datos son referenciales al 2017.

(Telégrafo, 2017).

Cifra que se ha visto afectada, ya que los países creadores de café como materia prima en medio oriente han comenzado a construir sus propias fábricas de café soluble, lo que hace que el Ecuador necesite mejorar y agilizar sus procesos en un mercado tan competitivo como este, lo que provocado un decrecimiento en las exportaciones hasta el año 2019.



(*Envíos de Café, Con La Peor Cifra Desde 2013 En Ecuador | El Comercio*, 2019.)

Este trabajo tiene una gran utilidad en el campo tecnológico e industrial, la misma que puede ser usada para estudios sobre la minería de datos y modelos predictivos así mismo para la mejora de este tipo de proyectos.

HIPÓTESIS O PREGUNTA DE INVESTIGACIÓN

La correcta toma de decisiones beneficia el alto rendimiento de la producción y minimiza errores en la asignación de recursos.

VARIABLES DE LA INVESTIGACIÓN

- **Variables independientes**: Café Tostado, Color del café, Temperatura del extractor, Temperatura de la centrifuga, Efecto del evaporador, Brix de Spray.

- **Variables dependientes**: Café Verde.

CAPÍTULO II
MARCO TEÓRICO

El Marco Teórico permite conocer todos los fundamentos que serán de utilidad en el contexto del proyecto. A continuación, se muestra el sustento teórico que permitirá tener una visión clara de los conceptos y teorías que son parte del proyecto de investigación.

Elementos teóricos y conceptuales

La minería de datos o minería de datos es un conjunto de técnicas que permiten la navegación automática o semiautomática de grandes bases de datos, con la finalidad de encontrar patrones, tendencias o reglas repetitivas que puedan explicar el comportamiento de los datos en un entorno determinado. (S.L., 2018). Básicamente es un proceso de la inteligencia de artificial y de inteligencia de negocios basada en la extracción de los datos. (Roşca & Rădoiu, 2015). Esta es una actividad en constante expansión aplicada a más y más disciplinas que han visto la utilidad de los datos de investigación para apoyar la toma de decisiones. En el campo de la inteligencia empresarial, ha avanzado enormemente en aquellas áreas relacionadas con la economía. (*Minería de datos en bibliotecas*, 2006).

Con el pasar del tiempo la minería de datos se llegado a convertir en una disciplina que es transparente para el mundo, debido a que la mayor parte del tiempo, no notamos que está pasando, pero nos encontramos generando información, cada vez que nos registramos para obtener una tarjeta de compras para una tienda de alimentos registramos nuestra tarjeta de crédito o navegamos en la web, estamos creando datos.(*Data Mining for the Masses*, 2012a) .

Esta se encarga de preparar, sondear y explorar los datos para sacar la información oculta y útil en ellos. Si los datos son leídos y analizados, pueden proporcionar, en conjunto, un verdadero conocimiento (futuras tendencias y comportamientos) que ayude en la toma de decisiones, ya que, para el responsable de un sistema, los datos en sí no son lo más relevante, sino la información que se encierra en sus relaciones, fluctuaciones y dependencias. Las bases de la minería de datos se encuentran en:

- La inteligencia artificial.
- El análisis estadístico.
- La Computación Gráfica.
- Las Bases de Datos.
- El Procesamiento Masivo.

(Yolanda Belinchón Monjas, 2019)

La Inteligencia Artificial hace posible que las máquinas aprendan de la experiencia, se ajusten a nuevas aportaciones y realicen tareas como seres humanos. La mayoría de los ejemplos de inteligencia artificial que escuchas hoy, desde computadoras que juegan al ajedrez hasta autos sin conductor, dependen en gran medida del aprendizaje profundo y el procesamiento del lenguaje natural. Con estas técnicas, las computadoras pueden capacitarse para realizar tareas específicas procesando grandes cantidades de datos e identificando patrones en los datos. (*Inteligencia artificial – Qué es y por qué es importante*, 2020).

Análisis Estadístico es la ciencia de recopilar, explorar y presentar grandes cantidades de datos para revelar patrones y tendencias potenciales. Las estadísticas se utilizan a diario en la investigación, la industria y el gobierno. (*Análisis estadístico ¿Qué es?*, 2020).

La Computación Gráfica es el campo de la informática visual, donde se utilizan computadoras para generar imágenes visuales y espaciales del mundo real. También podemos definirlo como el arte de transmitir información usando imágenes que son generadas mediante la computación (*Computación Gráfica - EcuRed*, 2019.).

Las Bases de Datos es una colección organizada de información estructurada, o datos, típicamente almacenados electrónicamente en un sistema de computadora. Una base de datos es usualmente controlada por un sistema de gestión de base de datos (DBMS). En conjunto, los datos y el DBMS, junto con las aplicaciones que están asociados con ellos, se conocen como un sistema de base de datos, que a menudo se reducen a solo base de datos. (*¿Qué es una base de datos?*, 2020).

El Procesamiento Masivo es la acumulación y manipulación de elementos de datos para producir información significativa. (*¿Qué es el Procesamiento de Datos?*, 2017).

Mediante el uso de minería de datos se pueden resolver los problemas de predicción, clasificación y segmentación. Es por eso que la mayoría de los datos recopilados, creados y administrados por las empresas hoy en día no están estructurados y son difíciles de manejar documentos de procesamiento de textos, hojas de cálculo, imágenes y Video, por lo que es necesario procesarlo automáticamente .(Timaran-Pereira et al., 2017).

Proceso de descubrimiento eficiente de patrones, desconocidos a priori, en grandes bases de datos. ([Agrawal and Shafer, 1996])

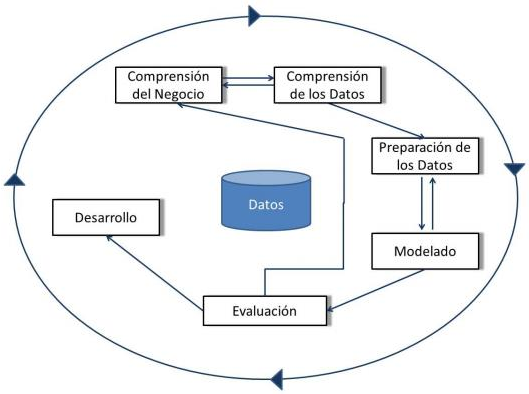


Figura 2: Proceso de Minería de Datos ([ZenTut, 2015])

(Maria Consuelo Justicia de la Torre, 2017)

Para descubrir patrones de relaciones útiles en un conjunto de datos se empezaron a utilizar métodos que fueron denominados de diferente forma. El término data mining, en inglés, no era, al principio, del agrado de muchos estadísticos, porque sus investigaciones estaban dirigidas a procesar y reprocesar suficientemente los datos, hasta que confirmasen o refutasen las hipótesis planteadas. (Febles Rodríguez & González Pérez, 2002) .

Por tanto, se plantean las siguientes preguntas: ¿Por qué es importante utilizar la minería de datos? La respuesta a esta pregunta es que

debido a que hay una gran cantidad de datos, se necesitan herramientas de análisis poderosas, porque algunas personas dicen "tenemos una gran cantidad de datos, pero carecemos de información". Con el rápido crecimiento, decenas de datos se recopilan y almacenan continuamente en enormes En el repositorio, está mucho más allá del entendimiento humano.(Yolanda Belinchón Monjas, 2019).

La etapa principal del proceso de minería de datos es el descubrimiento de reglas, que mostrarán nuevas relaciones entre variables o anomalías, según la empresa que utilice este proceso. Puede suceder que algunas reglas descubiertas no se puedan cambiar, sino que solo se pueden modificar para mejorar su rendimiento. Una vez que se encuentra una regla importante, se puede utilizar para estimar algunas variables de salida. En esta tecnología, la tecnología estadística tradicional y la tecnología de inteligencia artificial se complementan.

(Yolanda Belinchón Monjas, 2019)

El resultado del almacenamiento de datos se convierte en una "tumba de datos" Debido al gran número, rara vez se accede a estos archivos. Como resultado, las decisiones importantes no pueden basarse en datos previamente almacenados, sino intuitivas, porque no se basan en herramientas que extraigan el contenido más valioso en la cantidad de datos que existen. Desafortunadamente, estos procedimientos de toma de

decisiones son propensos a desviaciones y errores costosos y que requieren mucho tiempo. (Jiawei Han, 2006b).

Componentes de la minería de datos

Las componentes básicas de los métodos de la minería de datos son:

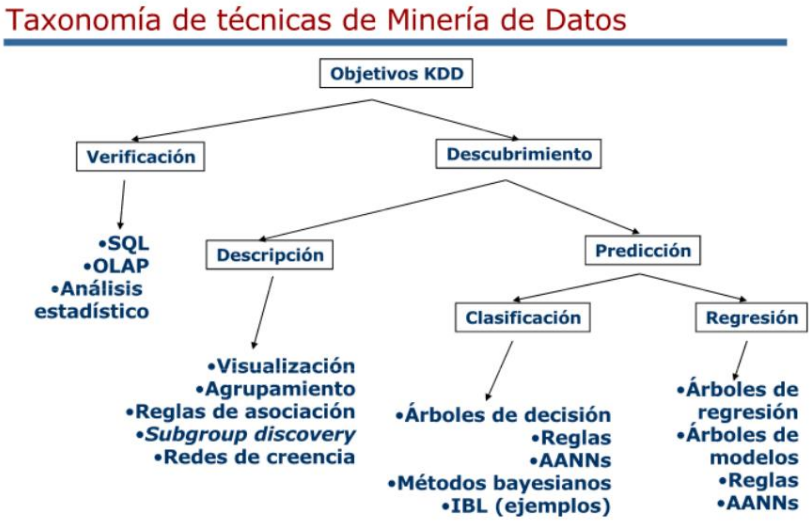
1. Lenguaje de representación del modelo: comprende las suposiciones y restricciones utilizadas en la representación empleada.
2. Evaluación del modelo: incluye el uso de técnicas de validación cruzada para la productividad y aplicación de principios como el de máxima verosimilitud o el de descripción mínima para evaluar la calidad descriptiva del modelo.
3. Método de búsqueda: puede dividirse en búsqueda de parámetros y del modelo, determinan los criterios que se siguen para encontrar los modelos.

Algunas de las técnicas más comunes usadas en la minería de datos son:

- Árboles de decisión y reglas de clasificación.
- Métodos de clasificación y regresiones no-lineales.
- Métodos basados en ejemplos prototípicos.
- Modelos gráficos de dependencias probabilísticas.
- Modelos relacionales.

(Febles Rodríguez & González Pérez, 2002)

Taxonomía de la Técnica de minería de datos.



(Febles Rodríguez & González Pérez, 2002)

Existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera. En (Witten and Frank 2005) se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semiautomático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Las bases de datos sobre las cuales la minería de datos trabaja son:

- Bases de datos relacionales.
- Bases de datos documentales

Bases de datos relacionales es una colección de tablas. Cada tabla consta de un conjunto de atributos como columnas o campos y puede contener un gran número de registros. Cada tupla representa un objeto, el cual se describe a través de los valores de sus atributos y se caracteriza por poseer una clave única o primaria que lo identifica.

Aunque las bases de datos relacionales son la fuente de datos para la mayoría de aplicaciones de minería de datos, muchas técnicas de minería de datos no son capaces de trabajar con toda la base de datos, sino que

sólo son capaces de tratar con una sola tabla a la vez. Lógicamente, mediante una consulta (por ejemplo, en SQL, en una base de datos relacional tradicional), podemos combinar en una sola tabla o vista minable aquella información de varias tablas que requiramos para cada tarea concreta de minería de datos. Por tanto, la presentación tabular, también llamada atributo-valor, es la más utilizada por las técnicas de minería de datos.

Bases de datos documentales Contienen descripciones de objetos como documentos de texto, que van desde simples palabras clave hasta resúmenes. Estas bases de datos pueden contener documentos no estructurados, documentos semiestructurados o documentos estructurados. La tecnología de minería de datos se puede utilizar para obtener asociaciones entre contenido, agrupar o clasificar objetos de texto.

Tipos de modelos de Minería de Datos

Modelos predictivos: Están diseñados para usar otras variables o campos en la base de datos para estimar el valor futuro o desconocido de la variable objetivo, que llamamos variables independientes o predictores. Por ejemplo, un modelo predictivo puede permitir estimar la demanda de nuevos productos según la inversión publicitaria.

Modelos descriptivos: Determinan los patrones utilizados para interpretar o agregar datos, es decir, se utilizan para explorar los atributos de los datos examinados, en lugar de predecir nuevos datos. Por ejemplo, una agencia de viajes quiere identificar grupos de personas con un mismo gusto para que puedan organizar diferentes ofertas para cada grupo para que puedan enviar esta información; para ello, analiza los viajes realizados por los clientes e infiere la descripción descriptiva de estos grupos. modelo.
¿A qué tipo de datos puede aplicarse la minería de datos? En principio, ésta puede aplicarse a cualquier tipo de información, siendo las técnicas de minería diferentes para cada una de ellas. Cuando el dato es textual se ha investigado en una nueva corriente que se le ha dado en llamar Minería de Texto.

Minería de Texto

El interés y la investigación sobre la Minería de Texto han aumentado, definiéndose como el proceso de extracción de información y conocimiento de los textos. La Minería de Texto analiza documentos. De modo más formal puede definirse del siguiente modo: "La minería de textos es el proceso de recopilar, organizar y analizar una gran cantidad de documentos. Tiene como objetivo brindar a los analistas y gerentes de empresas información sobre temas específicos. Esta información es muy útil para tomar decisiones. Decisiones, descubrimiento La relación entre diferentes hechos.

La minería de texto requiere también la previa preparación y almacenaje de los documentos o texto seleccionado. Se propone tareas tales como identificar los temas dominantes en un documento, elaborar índices de documentos, resumir textos de forma automática, clasificar los documentos, etc. Para realizarlas se han desarrollado distintas herramientas.

Minería Tecnológica

La minería tecnológica es la observación de la tecnología para detectar y analizar los cambios tecnológicos. Esto es, analizar las direcciones de lo que está ocurriendo ahora y basado en esto que ocurrirá en el futuro considerando el desarrollo de una tecnología en particular. Para realizar esto habrá que compilar y analizar información desde múltiples recursos. (Porter and Cunningham 2005) La tabla 2 muestra varios tipos de análisis tecnológicos que pueden ser realizados utilizando técnicas de Minería tecnológica.

Tabla 2. Posibles análisis tecnológicos a realizar utilizando Minería Tecnológica.

A	Vigilancia Tecnológica	Cataloga, caracteriza e interpreta las actividades de desarrollo tecnológico
B	Inteligencia Tecnológica Competitiva	Encontrar del ambiente externo ¿Quién está haciendo qué?
C	Previsión Tecnológica	Anticiparse a posibles desarrollos futuros en tecnologías particulares
D	Mapeo Tecnológico	Seguir los pasos de la evolución dentro de tecnologías relacionadas y familias de productos.
E	Evaluación Tecnológica	Anticiparse a posibles inentendidos, indirectas y consecuencias fuera de tiempo de un cambio tecnológico en particular.
F	Gestión de los procesos tecnológicos	Brindar información acerca de las tecnologías a los que toman decisiones

(Infante et al., 2019)

Algoritmos de minería de datos.

Los algoritmos de modelos predictivos facilitan al proceso de la toma de decisiones las cuales pueden ser tomadas a través de varias técnicas que se pueden aplicar como estas tenemos:

- Árbol de decisiones.
- Redes neuronales.
- Máquinas de vectores.
- Análisis Bayesiano.
- Regresión Logística.
- Regresión Lineal.
- K-Vecinos más Cercanos.
- Series Temporales y Data Mining.
- Ensemble Models.
- Potenciación del Gradiente.
- Modelos de Respuesta Incremental.

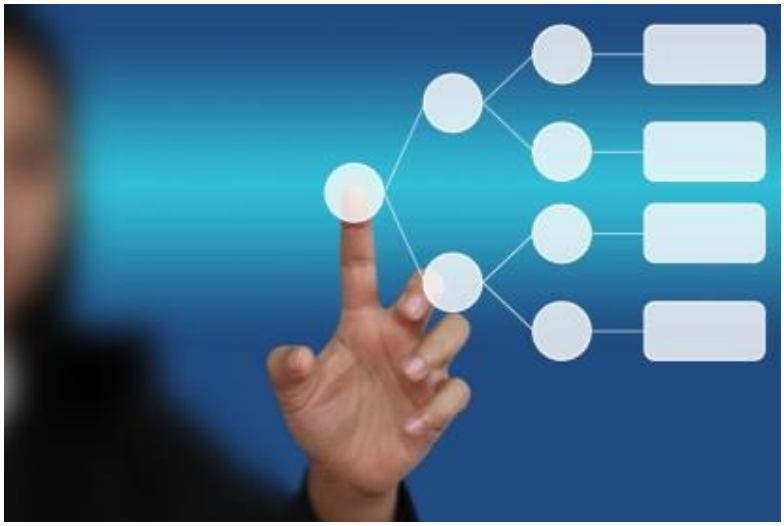
Algoritmo EM: Este algoritmo define parámetros analizando los datos y predice la posibilidad de una salida futura o evento aleatorio dentro de los parámetros de datos. Por ejemplo, el algoritmo EM podría intentar predecir el momento de una siguiente erupción de un géiser según los datos de tiempo de erupciones pasadas.

Algoritmo CART: "CART" es una sigla en inglés que significa análisis de árbol regresivo y de clasificación. Al igual que los análisis de árboles de decisión, organiza los datos según opciones que compiten, como si una persona ha sobrevivido a un terremoto. Al contrario que los algoritmos de árboles de decisión, que sólo pueden clasificar una salida o una salida numérica basada en la regresión, el algoritmo CART puede usar los dos para predecir la probabilidad de un evento.

Árbol de decisión: Los algoritmos de árbol de decisión consisten en organizar los datos en elecciones que compiten formando ramas de influencia después de una decisión inicial. El tronco del árbol representa la decisión inicial, y empieza con una pregunta de sí o no, como tomar o no el desayuno. Tomar desayuno y no tomar desayuno serían las dos ramas

divergentes del árbol, y cada elección posterior tendría sus propias ramas divergentes que llevan a un punto final.

Son **modelos de clasificación muy utilizados** que tratan de encontrar la variable que permita dividir el dataset en grupos lógicos que son más diferentes entre sí. Cada árbol se va descomponiendo en distintas ramas y hojas que representan cada clasificación en función de las condiciones que se van seleccionando hasta llegar a la resolución del problema. Estos modelos son de gran ayuda a la hora de determinar las decisiones a lo largo de un proceso como por ejemplo el funnel de compra.



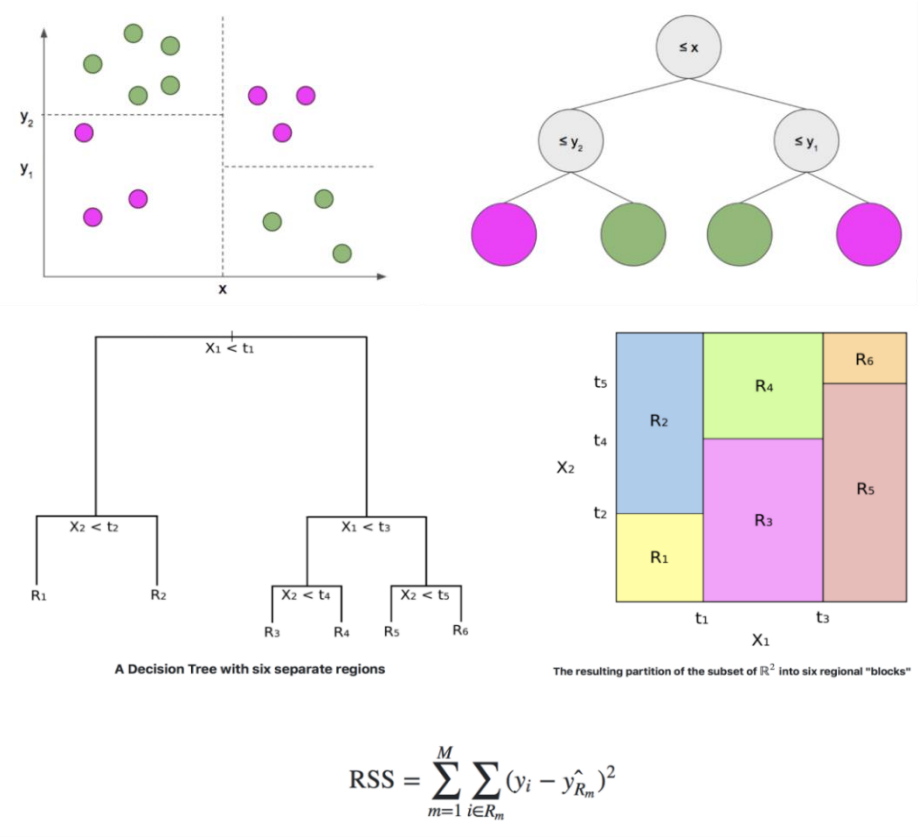
(*Arboles de Decision y Random Forest*, 2018)

- Fácil de entender.
- Útil en exploración de datos: identificar importancia de variables a partir de cientos de variables.
- Menos limpieza de datos: outliers y valores faltantes no influencian el modelo (A un cierto grado).
- El tipo de datos no es una restricción.
- Es un método no paramétrico (i.e., no hay suposición acerca del espacio de distribución y la estructura del clasificador).

(*Arboles de Decision y Random Forest*, 2018)

- El enfoque *classification and regression tree (CART)* fue desarrollado por Breiman et al. (1984).
- Son un tipo de algoritmos de aprendizaje supervisado.
- Principalmente usados en problemas de clasificación.

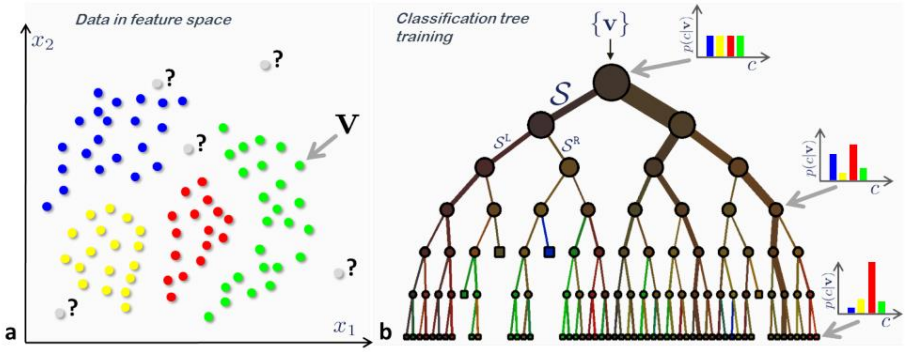
- Las variables de entrada y salida pueden ser categóricas o continuas.
- Divide el espacio de predictores (variables independientes) en regiones distintas y no sobrepuestas.



$$\text{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y_{R_m}})^2$$

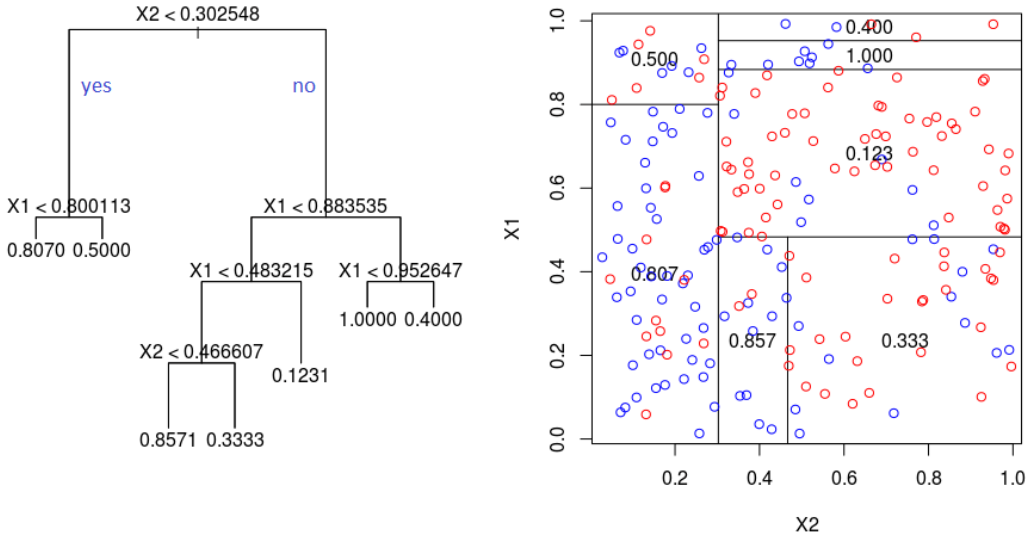
(*Arboles de Decision y Random Forest*, 2018)

- Se divide la población o muestra en conjuntos homogéneos basados en la variable de entrada más significativa.
- La construcción del árbol sigue un enfoque de división binaria recursiva (top-down greddy approach). Greedy -> analiza la mejor variable para ramificación sólo en el proceso de división actual.



Un árbol de regresión consiste en hacer preguntas de tipo ¿xk≤cxk≤c? para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiper-rectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado ^yy^.

En la siguiente figura se ilustra el árbol en el lado izquierdo y la partición del espacio en el lado derecho. La partición del espacio se hace de manera repetitiva para encontrar las variables y los valores de corte cc de tal manera que se minimice la función de costos ∑i=ni=1(yi−^yi)2∑i=1i=n(yi−y^i)2.

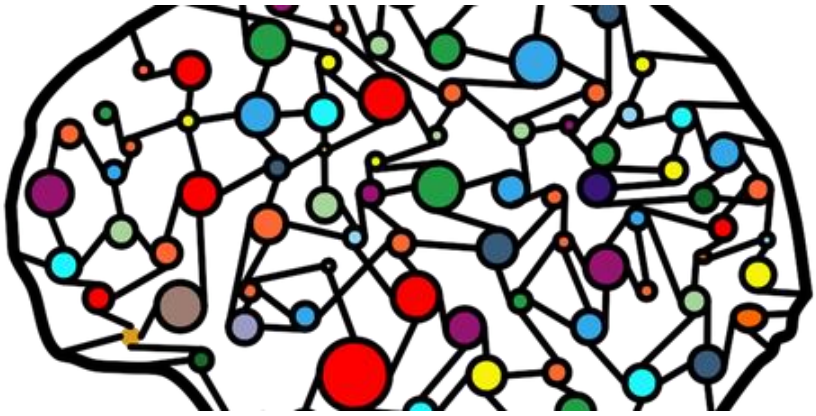


(Hernández, 2020.)

Redes Neuronales

La **Inteligencia Artificial y el Deep Learning** han puesto muy de moda esta técnica tan sofisticada de reconocimiento de patrones que imita las neuronas del cerebro humano ya que es capaz de modelar relaciones extremadamente

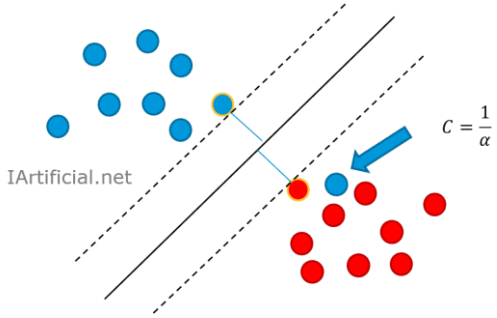
complejas y suele utilizarse cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y los de salida.



(*¿Sabes En Qué Se Diferencian Las Redes Neuronales Del Deep Learning?*, 2019)

Máquinas de Vectores de Soporte (SVM)

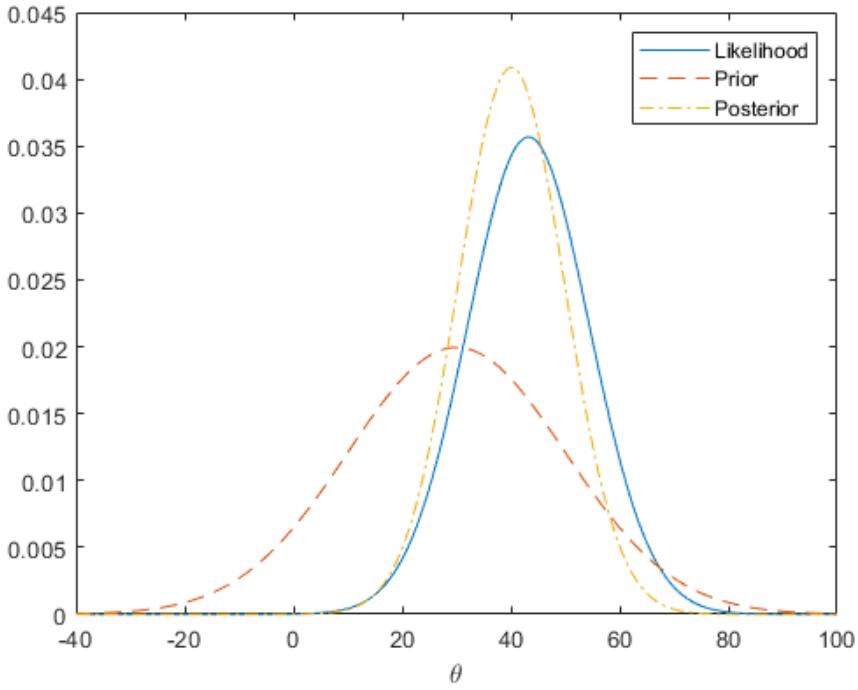
Son algoritmos de aprendizaje automático supervisado de cara a reconocer patrones, estando relacionados con problemas de clasificación o regresión.



(Heras, 2020)

Análisis Bayesiano

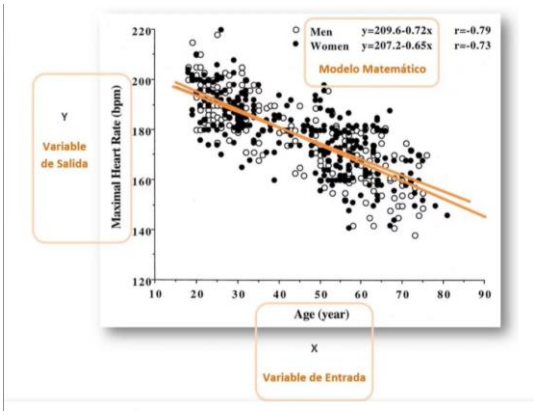
Se trata de una inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta.



(*Análisis Bayesiano Para Un Modelo de Regresión Logística - MATLAB & Simulink Example - MathWorks América Latina*, 2019)

Regresión Logística

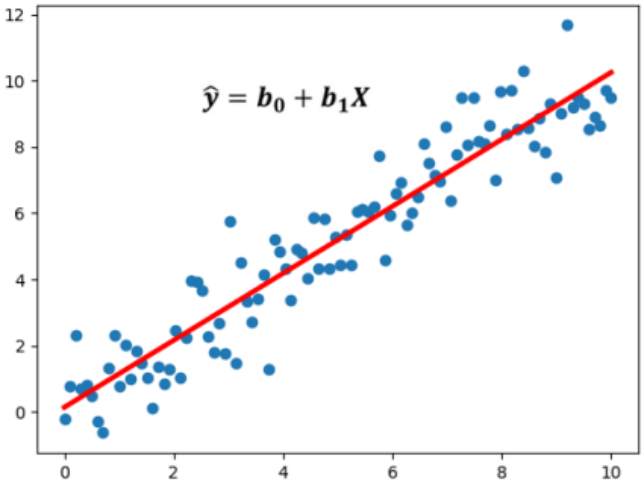
Las regresiones logísticas son utilizadas para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictivas. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. Por ejemplo, puede utilizarse para predecir el riesgo crediticio.



("Qué es y cómo interpretar una regresión logística - Incluye Ejemplo," 2019)

Regresión Lineal

La regresión lineal consiste en una línea recta que muestra el “mejor encaje” de todos los puntos de los valores numéricos. También se llama el método de los mínimos cuadrados porque calcula la suma de las distancias al cuadrado entre los puntos que representan los datos y los puntos de la línea que genera el modelo. Así, la mejor estimación será la que minimice estas distancias.



Ejemplo de una Regresión Lineal Simple

Series Temporales y Data Mining

Este método combina una mezcla de técnicas de data mining tradicional como sampleado, clustering y árboles de decisión, con otras de forecasting con el fin de mejorar las predicciones sobre datos recopilados como ventas por meses o trimestres, llamadas por día, o visitas a nuestra web por hora.

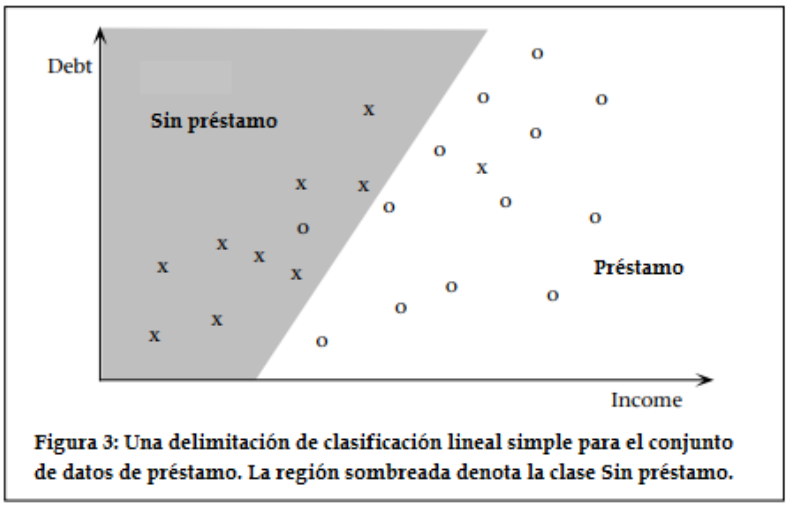
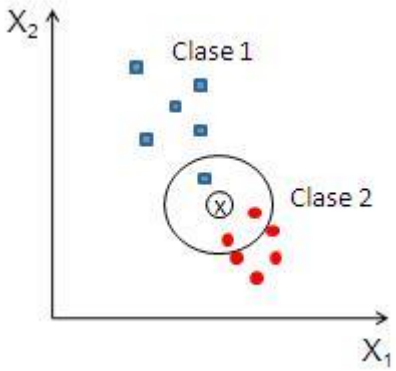


Figura 3: Una delimitación de clasificación lineal simple para el conjunto de datos de préstamo. La región sombreada denota la clase Sin préstamo.

(“Los Métodos Del Data Mining o Minería de Datos,” 2019.)

K-Vecinos más Cercanos

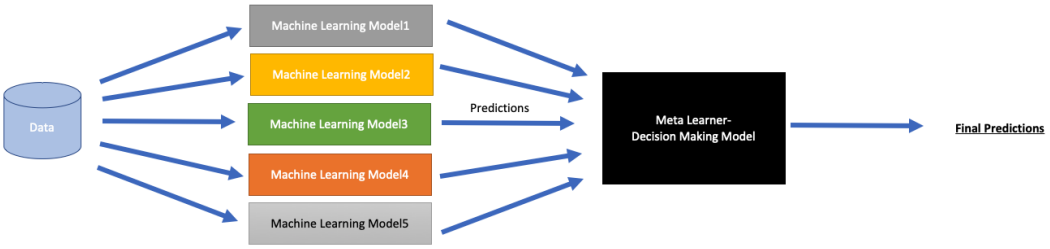
La frase “Dime con quién vas y te diré quién eres” nos explica a la perfección cómo funciona este algoritmo de agrupamiento o clustering. Consiste en reconocer patrones para conocer la probabilidad de que un elemento pertenezca a una clase según su cercanía en el espacio a los elementos de esa clasificación.



(*Regla de Los K Vecinos Más Cercanos - EcuRed*, 2018.)

Ensemble Models

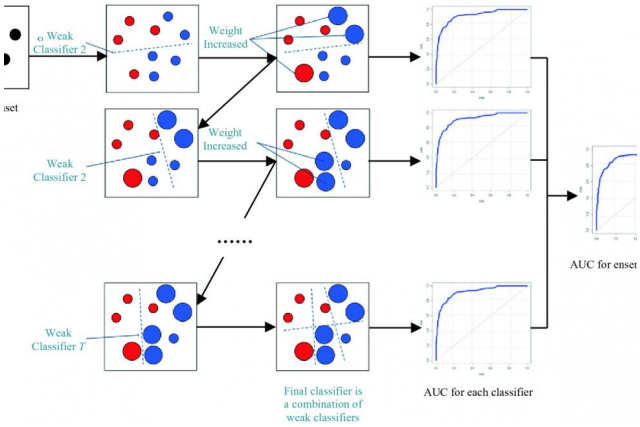
Es famoso por su precisión debido a la disponibilidad de algoritmos de boosting y bagging. Crea un nuevo modelo entrenando varios modelos similares combinando los resultados para mejorar la precisión, reducir la varianza y los sesgos e identificar el mejor modelo para usar con nuevos datos.



(Tyagi, 2020)

Potenciación del Gradiente

Lleva a cabo un resampling de nuestro dataset para generar unos resultados que formen una media ponderada del conjunto de datos.



(“Impulso de gradiente - Lo que necesitas saber — Aprendizaje automático,” 2020)

Modelos de Respuesta Incremental

Suele utilizarse para reducir el Churn o comprobar la efectividad de diferentes acciones de Marketing. Se modela el cambio de probabilidad causado por una acción.

(Spain, 2020b)

Las herramientas de minería de datos realizan análisis y pueden descubrir patrones de datos importantes, lo que contribuye en gran medida al negocio, ya que la brecha es cada vez mayor entre datos e información que requiere un desarrollo sistemático de herramientas de minería de datos que convertirá las tumbas de datos en “pepitas de oro” de conocimiento.

(Jiawei Han, 2006a)

Metodologias de mineria de datos

Existen varias metodologias de mineria de datos entre las cuales tenemos:

- Semma.
- Catalyst.
- CRISP-DM.

La metodología **SEMMA** se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Se recomienda especialmente usarlo con

el software de minería de datos de SAS. El producto organiza sus herramientas llamadas "nodos" según las diferentes etapas que componen el método. En otras palabras, el software proporciona un conjunto de herramientas especiales para la etapa de muestreo, otras herramientas especiales para la etapa de exploración, etc. Sin embargo, los usuarios pueden usarlo de acuerdo con cualquier otro método de minería de datos. (como CRISP-DM por ejemplo).

La metodología **Catalyst**, en sus dos modelos, está compuesta por una serie de pasos llamados “boxes”. El concepto es que luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles.

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. La sucesión de fases, no es necesariamente rígida. Cada fase es descompuesta en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, pero en ningún momento se propone como realizarlas. Es decir, CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto, pero no especifica cómo llevarlas a cabo.

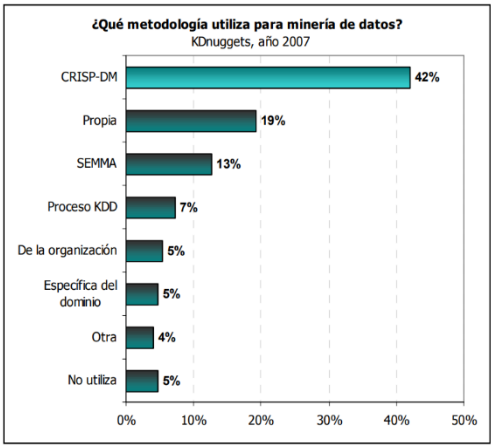


Fig. 1. Encuesta realizada por la KDnuggets en el año 2007

(Ing. Juan Miguel Moine, 2019)

Herramientas de desarrollo para minería de datos

R: Es un lenguaje de programación con un enfoque analítico, lanzado en 1995 como mejora del lenguaje S. Escrito en C, Fortran y en sí mismo, el proyecto cuenta actualmente con el apoyo de la R Foundation for Statistical Computing. Este lenguaje proporciona un amplio abanico de herramientas estadísticas, como modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento.

Ventajas

- Excelente gama de paquetes de código abierto y de alta calidad. R tiene un paquete para casi todas las aplicaciones cuantitativas y estadísticas imaginables. Esto incluye redes neuronales, regresión no lineal, filogenia, cartografía, mapas y muchos, muchos otros.
- La instalación básica viene con funciones y métodos estadísticos integrales muy completos. R también maneja el álgebra de matriz particularmente bien.
- La visualización de datos es una fortaleza clave con el uso de bibliotecas como ggplot2.

Desventajas

- R no es un lenguaje rápido. Esto no es un accidente. R fue diseñado a propósito para facilitarle el análisis de datos y las estadísticas. No fue diseñado para ser rápido.

- R es lento en comparación con otros lenguajes de programación, para la mayoría de los propósitos, es lo suficientemente rápido.
- Especificidad de dominio de R es fantástico para fines estadísticos y científicos de datos. Pero no es tan fantástico para programaciones de propósito general.
- R no soporta gráficos en tres dimensiones o dinámicos. El resultado de cada informe puede ser algo pobre visualmente y bastante anticuado en comparación con el ofrecido por otros programas.

(Morales, 2019)

En cuanto a Machine Learning, R tiene implementados una gran cantidad de algoritmos, como consecuencia de las diferentes líneas de investigación de grupos que dieron pie a su creación, debido precisamente al hecho de que R nació en el ámbito académico.

Como contra en R, está su curva de aprendizaje, que suele ser más lenta y complicada si la comparamos con la de Python.

(*Python o R. ¿Qué lenguaje utilizar para el análisis de datos?*, 2016)

Python: Es un lenguaje de programación multiparadigma, por lo que puede soportar parcialmente la orientación a objetos, programación imperativa y programación funcional. La historia de Python parte a partir de 1991 desde que Guido van Rossum lo presentó y desde entonces, se ha convertido en un lenguaje de uso general extremadamente popular, y se utiliza impliamente en la comunidad de data science.

Ventajas

- Python es un lenguaje de programación de uso general muy popular y general.
- Es muy flexible y open source.
- Es rápido en su ejecución en comparación a otros lenguajes como R o Java.
- Python es un lenguaje fácil de aprender. La baja barrera de entrada lo convierte en un primer idioma, lo que es ideal para aquellos que son nuevos en programación.

- Permite resolver problemas de machine learning.
- Consta con paquetes como pandas, scikit-learn y Tensorflow hacen de Python una opción sólida para aplicaciones avanzadas de aprendizaje automático.
- Las numerosas librerías creadas para esta finalidad como: Numpy y Pandas que implementan funciones para cálculos matemáticos y estadísticos, Mlpy con algoritmos de aprendizaje máquina, Matploit que permite la visualización y representación gráfica de los datos.
- Su integración con aplicaciones como MongoDB, Hadoop o Pentaho.
- Si a todo ello se le suma la fácil y rápida curva de aprendizaje junto con su versatilidad, hacen de Python un lenguaje de gran calidad para los analistas de datos.

Contras

- Python es un lenguaje de tipo dinámico, lo que significa que debemos ser muy cuidadosos. Los errores de tipo como pasar una string como un argumento a un método que espera un número entero deben esperarse de vez en cuando.
- Para las simulaciones físicas el lenguaje Python puede resultar complejo, ya que no trabaja con matrices por defecto, tal como ocurre con otros lenguajes como Matlab. En definitiva, este lenguaje te es útil siempre que no dependas de una matriz o tengas que trabajar con un vector complejo, ya que de lo contrario debes importar bibliotecas.

(Morales, 2019)

Los dos motivos principales que han generado un creciente uso de Python en el campo del análisis de datos son:

(*Python o R. ¿Qué lenguaje utilizar para el análisis de datos?*, 2016)

SQL: Lenguaje de consulta estructurado define, administra y consulta bases de datos relacionales. El lenguaje apareció en 1974 y desde entonces ha sufrido muchas implementaciones, pero los principios básicos siguen siendo los mismos.

Ventajas

- Muy eficiente en consultas, actualización y manipulación de bases de datos relacionales.
- La sintaxis declarativa hace de SQL un lenguaje muy legible. No hay ambigüedad sobre lo que se debe hacer.
- SQL utilizado en una amplia gama de aplicaciones, por lo que es un lenguaje muy útil para estar familiarizado.
- Los módulos como SQLAlchemy hacen que la integración de SQL con otros lenguajes sea sencillo.

(Morales, 2019)

Contras

- Las capacidades analíticas de SQL son bastante limitadas: más allá de agregar y sumar, contar y promediar datos, sus opciones son limitadas.
- Para los programadores que vienen de un contexto imperativo, la sintaxis declarativa de SQL puede presentar una curva de aprendizaje.
- Hay muchas implementaciones de SQL como PostgreSQL, SQLite, MariaDB. Todas son lo suficientemente diferentes como para hacer que la interoperabilidad sea un dolor de cabeza.

(Morales, 2019)

JAVA: Java es un lenguaje extremadamente popular que se ejecuta en la Máquina Virtual Java. Es un sistema informático abstracto que permite una portabilidad perfecta entre plataformas. Actualmente respaldado por Oracle Corporation**.**

Ventajas

- Se puede aplicar en muchos sistemas y aplicaciones modernas se basan en un back-end de Java.
- Tiene la capacidad de integrar métodos de ciencia de datos directamente en la base de código existente es poderosa.

- Es un buen lenguaje cuando se trata de garantizar la seguridad de tipos.
- Se puede aplicar en aplicaciones de big data de misión crítica.
- Java es un lenguaje compilado de propósito general y alto rendimiento. Lo que lo hace adecuado para escribir eficientes códigos de producción ETL y algoritmos de machine learning muy intensivos computacionalmente.

Contras

- Para análisis ad-hoc y aplicaciones estadísticas más dedicadas, la verbosidad de Java hace que sea una primera opción poco probable.
- Los lenguajes de script de tipado dinámico como R y Python se prestan a una productividad mucho mayor.
- En comparación con los lenguajes específicos de dominio como R, no dispone de muchas librerías disponibles para métodos estadísticos avanzados.

(Morales, 2019)

Scala: Desarrollado por Martin Odersky y lanzado en 2004, Scala es un lenguaje que se ejecuta en la Máquina Virtual Java. Es un lenguaje de múlti-paradigmático, que permite tanto enfoques orientados a objetos como funcionales. El framework de computación de cluster Apache Spark está escrito en Scala.

Ventajas

- Scala + Spark = Computación en clúster de alto rendimiento.
- Scala es un lenguaje ideal para quienes trabajan con conjuntos de datos de gran volumen.
- Scala pueden tener lo mejor de la programación orientada a objetos como funcional.
- Scala se compila en el bytecode de Java y se ejecuta en una JVM. Esto le permite la interoperabilidad con el lenguaje Java en sí, haciendo de Scala un lenguaje de propósito general muy poderoso, además de ser adecuado para la ciencia de datos.

Contras

- Scala no es un lenguaje sencillo para comenzar a utilizar si está empezando. Lo mejor es descargar sbt y configurar un IDE como Eclipse o IntelliJ con un complemento específico de Scala.
- La sintaxis y el sistema de tipos se describen con frecuencia como complejos. Esto hace que la curva de aprendizaje sea pronunciada para aquellos que vienen de lenguajes dinámicos como Python.

(Morales, 2019)

Julia: Lanzada en 2011, Julia impresionó al mundo de la computación numérica. Su perfil se elevó gracias a la adopción temprana por parte de varias organizaciones importantes, incluidas muchas de la industria financiera.

Ventajas

- Julia es un lenguaje compilado Just-In-Time, que le permite ofrecer un buen rendimiento. También ofrece las capacidades de simplicidad, tipado dinámico y scripting de un lenguaje interpretado como Python.
- Julia fue diseñada específicamente para el análisis numérico. Pero también ofrece programación de propósitos generales.
- Legibilidad. Muchos usuarios del lenguaje mencionan ésto como una ventaja clave.

Contras

- Madurez. Como nuevo idioma, algunos usuarios de Julia han experimentado inestabilidad al usar paquetes complementarios. Pero el núcleo del lenguaje es, al parecer, lo suficientemente estable para usar en producción.
- Los paquetes limitados son otra consecuencia de la juventud del lenguaje y de la pequeña comunidad de desarrollo. A diferencia de R y Python, Julia no tiene la posibilidad de disponer de paquetes.

(Morales, 2019)

Matlab: es un lenguaje de computación numérica que se utiliza en el mundo académico y en la industria. Desarrollado y licenciado por MathWorks, una compañía establecida en 1984 para comercializar el software.

Ventajas

- Diseñado para la computación numérica. MATLAB es adecuado para aplicaciones cuantitativas con requisitos matemáticos sofisticados, como procesamiento de señales, transformaciones Fourier, álgebra matricial y procesamiento de imágenes.
- Visualización de datos. MATLAB tiene incorporadas grandes capacidades de ploteado.
- MATLAB se enseña con frecuencia como parte de cursos de pregrado en asignaturas cuantitativas como Física, Ingeniería y Matemáticas Aplicadas. Como consecuencia, es ámpliamente utilizado en estos campos.

Contras

- Licencia propietaria. Dependiendo del caso (uso académico, personal o empresarial) es posible que tengamos que desembolsar una gran cantidad de dinero. Existen alternativas gratuitas disponibles como Octave.
- MATLAB no es una opción obvia para programación de propósito general.

(Morales, 2019)

R es una buena opción cuando el análisis de datos requiere una computación independiente o un análisis individual en los servidores, mientras que Python se puede usar cuando el análisis de datos requiera ser integrado con las aplicaciones webs o si se necesita incorporar el código de análisis estadístico en una base de datos de producción.

Se debe tener en cuenta, en función de nuestros conocimientos previos en programación, estadística, etc., si R o Python resultaran más fácil de aprender y poner en práctica, se debe tener en cuenta que el objetivo es resolver el problema que tenemos entre manos, y aprender el uso de las

herramientas para resolverlo no debe convertirse en el núcleo del problema en sí.

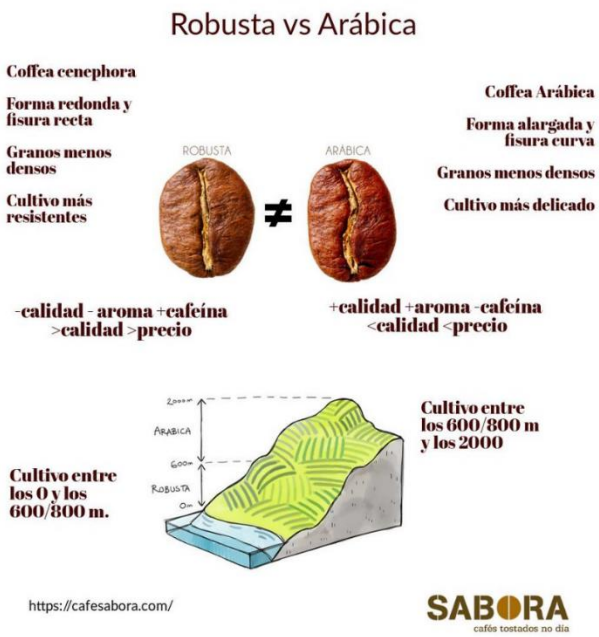
(*Python o R. ¿Qué lenguaje utilizar para el análisis de datos?*, 2016)

Origen del café

El café es el segundo producto más consumido a nivel mundial después del agua, este tiene sus orígenes en África. Es tan antiguo, que no existen documentos escritos sobre cuando se comenzó a tomar. Su nombre es una derivación de la palabra Qahwa, que significa excitante, energético, vigoroso.

Todo lo que tenemos son algunas leyendas tribales que han perdurado durante años, como sabiduría popular. La más famosa dice así: *"*Se cuenta que los miembros de las tribus observaron cómo las cabras presentaban un comportamiento más energético de lo normal, tras comer cierto tipo de bayas. Algunos de ellos decidieron probarla y comprobaron esos beneficios, siendo las primeras personas en consumir café en toda la historia de la humanidad.". A pesar de no haber documentos fidedignos sobre cuando se comenzó a consumir, se sabe que no fue hasta el siglo XV, cuando se comenzaron a tener las primeras evidencias del consumo del esta.

El café comenzó a venderse entre los monasterios sufíes de Etiopía y Yemen. Su expansión alcanzó la población islámica en el siglo XVI y saltó a Europa a principios del siglo XVII. Su comercio tuvo lugar entre la República de Venecia y el norte de África, en el mismo siglo llegó al continente asiático a través de la India, abriendo así mercados en Japón y China. Llegó al Nuevo Mundo a mediados del siglo XVII y llegó a Nueva York, pero no fue hasta 1773 que se convirtió en un borracho estadounidense. ("Historia del café - Descubre todo sobre sus orígenes," 2015)



(Sabora, 2016)

El café se encuentra en el centro de la cereza que está dentro de la planta del café, los granos son considerados como un tesoro escondido apreciado por personas de todo el mundo. Según la Asociación Británica del Café, en general, el Reino Unido debe alrededor de 95 millones de tazas de café al día, este se cultiva alrededor de más de 50 países alrededor del “cinturón del café", incluidos lugares como África, América Latina y Asia.

Café África

Se cree que el café africano es uno de los mejores del mundo debido a su sabor maravillosamente distintivo. El café tradicionalmente africano cuenta con las siguientes características:

- Almibarado.
- Acidez media.
- Tostado ligero a medio.

América Latina

Se cree que es una de las capitales cafeteras del mundo, el café de América Latina constituye la mayoría de las mezclas que se encuentran en los supermercados en la actualidad. El sabor se disfruta universalmente, lo que se atribuye principalmente a su sabor bien completo. Algunas de las características clave que posee el café latinoamericano son:

- Sabor a nuez.
- Baja acidez.
- Tostado ligero o medio.

Asia

Asia es el hogar de algunas de las mezclas de café más exclusivas con sabores intensos y únicos como ningún otro. Por lo general, puede esperar que el café asiático sea:

- Terroso.
- Suave acidez.
- Tostado oscuro.

(*La historia del café | Nescafe Argentina*, 2015)

El **café arábico** crece en las alturas y es más sensible a las temperaturas y los parásitos mientras que el café robusto es más resistente, crece también en lugares bajos con temperaturas que a veces superan los 30º

	Arabica	Robusta
Más cafeína		✓
Más aroma	✓	
Más cuerpo		✓
Más dulce	✓	
Más delicado	✓	
Más cremoso		✓

(passalacqua, 2019)

El Café Robusta tiene el doble de cafeína que el arábica. Es un tipo de variedad original de África centra que, al crecer en zonas secas, es poco digestivo, tiene un gusto final amargo, con mucho cuerpo y pocos perfumado. Su cultivo representa el 43% de la producción mundial y es un café más económico que la variedad (cafesaula, 2014),

Robusta es una especie diploide. Es un arbusto más grande que el cafeto de arábica y su crecimiento es robusto. El sistema de raíces de robusta, aunque grande es poco profundo comparado con arábica, y la masa

de raíces de alimentación está confinada a las capas superiores del suelo. Las hojas son anchas, grandes y de color verde pálido. Las flores son blancas y fragantes, y forman racimos mayores que los de arábica. Las flores se abren al séptimo u octavo día, después de recibir la lluvia. Al contrario del arábica, el robusta es auto estéril, es decir que su óvulo no puede ser fertilizado con su propio polen por lo que necesita una polinización cruzada. Las cerezas son pequeñas, pero más numerosas en cada nudo que en el arábica, variando de 40 a 60 o más. Maduran en unos 10 u 11 meses y están generalmente a punto para la recolección dos meses más tarde que las de arábica. (*11.9.1-Calidad Del Café-Robusta - La Especie*, n.d.)

Hanna dice: “Robusta se considera de calidad mucho más baja que Arábica. Sin embargo, es debatible qué porcentaje de este problema de calidad se debe atribuir a la genética, en comparación con el hecho que Robusta generalmente no tiene los mismos estándares de calidad que Arábica”. (Kanniah, 2020)

Café arábico: Arábica representa alrededor del 70% de la producción mundial de café, y generalmente, se considera que produce un café de mejor sabor. Por supuesto, el sabor exacto de un café depende en gran medida de su origen, método de procesamiento y otros factores.

Gonzalo Hernández es el presidente de Coffea diversa, una empresa de abastecimiento de café verde ubicada en Costa Rica. Él dice: “No hay una receta o descripción general en términos del perfil de sabor de Arábica, dependiendo de las variables. El perfil de sabor podría ser achocolatado, especiado, floral, acaramelado, con una acidez brillante, acidez seca o de baja acidez, jugoso, afrutado, etc.”. (Kanniah, 2020)

Tipos de tostado.

Tostado Ligeros: Un tostado ligero tiene un sabor de grano suave y tostado con un cuerpo liviano, una marcada acidez y sin aceite en la superficie de los granos. El tostado más ligero se llama canela clara.

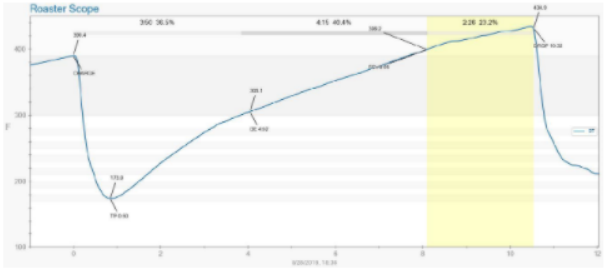
Tostados medios: Un tostado medio tendrá más cuerpo y menos acidez que un tostado ligero, pero tampoco tiene aceite en la superficie de los granos. Se conoce comúnmente como tostado americano.

Tostados intensos: Los tostados oscuros son los más fuertes y su sabor dulce ahumado también puede ser amargo. Los granos tostados franceses son casi negros. Los granos tostados italianos son negros, caramelizados y aceitosos.

(*La historia del café* | *Nescafe Argentina*, 2015)

“La temperatura del grano es crucial”, me cuenta Andrew. “Pero igualmente crucial es que debe usarse un reloj y un cronómetro. El tiempo y la temperatura lo son todo”.

También te puede gustar ¿Qué Ocurre Durante El Tueste Del Café? Los Cambios Químicos



Las lecturas de la temperatura del grano. Crédito: Zach Latimore

(Latimore, 2019)

Tabla 2. Etapas y cambios físicos en el proceso de torrefacción[14]

TEMPERATURA DEL GRANO (°C)	COLOR	VOLUMEN	PROCESO
100	Amarillo		Desecación y pérdida de agua.
120-130	Castaño		Reacciones de reducción de azúcares y aminoácidos
130-180	Marrón	Aumenta	Caramelización de azúcares.
180-200	Marrón	Aumenta	Producción de CO_2 por pirogenación de carbohidratos, proteínas y grasas
200-230	Marrón	Aumenta	Agrietamiento del grano (crepitación) y afloramiento del aceite en la superficie.
250	Negro, sin brillo	Deja de aumentar	Sobretorrefacción, se carboniza y el aroma desaparece.

Fuente: Extracción de aceite esencial a partir de café brocado.

(Duarte, 2002)

MARCO LEGAL.

El **Reglamento de Régimen Académico**, Capitulo III De la Estructura Curricular, artículo 21, Unidades de organización curricular en las carreras técnicas y tecnológicas superiores y equivalentes; y, de grado.

- **Unidad de Titulación:** Es la unidad curricular que incluye las asignaturas, cursos o sus equivalentes, que permiten la validación académica de los conocimientos, habilidades y desempeños adquiridos en la carrera para la resolución de problemas, dilemas o

> desafíos de una profesión. Su resultado final fundamental es: a) el desarrollo de un trabajo de titulación, basado en procesos de investigación e intervención o, b) la preparación y aprobación de un examen de grado de carácter complexivo (CES, 2017a, p.12).

En cuando al trabajo de titulación, el CES la define como:

> Propuesta innovadora que contenga, como mínimo, una investigación exploratoria y diagnóstica, base conceptual, conclusiones y fuentes de consulta. Para garantizar su rigor académico, el trabajo de titulación deberá guardar correspondencia con los aprendizajes adquiridos en la carrera y utilizar un nivel de argumentación coherente con las convenciones del campo del conocimiento. (p. 13).

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN

Metodología de la investigación es el elemento que enlaza el sujeto con el objeto de la investigación.(“Metodología de la investigación (definición y conceptos),” 2018). Sin ella es inadmisible llegar a la lógica que trasfiere al conocimiento científico. También es el conjunto de procedimientos y técnicas que se aplican de manera ordenada y sistemática en la realización de un estudio.(*Significado de Metodología de la investigación*, 2015) Para el trabajo investigativo debe acompañado de los debidos métodos y tipos de investigación, los cuales permiten orientar la búsqueda de conocimiento del proyecto además de las técnicas y herramientas de levantamiento de información pertinente al problema a resolver.

El tipo de investigación es por factibilidad que describe lo fácil o difícil que puede resultar hacer algo. Cuando se establece una meta en el trabajo, se piensa en la factibilidad a largo plazo de lograr lo que se desea. Indica si vale la pena invertir en un proyecto.

Cuando se habla sobre la factibilidad de un proyecto, realmente se está discutiendo si se puede lograr o no, ¿qué tan factible es? Por ejemplo, si se desea cuestionar el plan de un hombre de pintar toda su casa en un solo fin de semana, se le diría que debe analizar la factibilidad de esa tarea. Esto permitirá preguntar si se puede hacer o no.(Corvo, 2019)

Por lo expuesto en este capítulo que está dedicado a la identificación del tipo de investigación, población y herramientas utilizadas para el levantamiento de la información. Así mismo se realizará el análisis de la información obtenida.

TIPO DE INVESTIGACION

La metodología de la investigación utilizada en el trabajo de titulación para poder generar resultados del proceso de minería de datos, el cual tiene un enfoque cualitativo, el cual está centrado en la interpretación y los resultados descriptivos. (*Método Cualitativo - Concepto, Características y*

Ejemplos, 2019) . Produce información sólo en los casos particulares que estudia, por lo que es difícil generalizar, sólo se puede hacer mediante hipótesis.

Los métodos cualitativos se basan en principios teóricos como la fenomenología, la hermenéutica y la interacción social El método de recolección de información utilizado es diferente al método cuantitativo porque no puede reflejarse en cantidad. La idea es explorar las relaciones sociales y describir la realidad a medida que el protagonista la vive. (“¿Qué es el método cualitativo?,” 2015)

METODOLOGIA DE MINERIA DE DATOS

El tipo de investigación tomado en cuenta será una investigación por objetivos la cual se basa en la investigación pura y aplicada.

Existen diferentes metodologías dentro del campo de la Minería de datos, de acuerdo con lo investigado, se procedió a comparar las metodologías para seleccionar la más idónea para el trabajo de investigación es CRISP–DM, la cual es se establece en un proyecto como una secuencia de fases que son:

- Compresión del negocio.
- Comprensión de los datos.
- Preparación de los datos.
- Modelado.
- Evaluación.

Comprensión del negocio:

El objetivo de esta fase es alinear los objetivos del proyecto de data mining con los objetivos del negocio. Tratando así de evitar embarcarnos en un proyecto de minería de datos que no produzca ningún efecto real en la organización.

En esta fase deberemos ser capaces de:

- Establecer los objetivos de negocio.
- Evaluar la situación actual.

- Fijar los objetivos a nivel de minería de datos.
- Obtener un plan de proyecto.

Comprensión de datos:

Dos puntos clave en esta fase: conocer los datos, estructura y distribución, y la calidad de los mismos.

En esta fase deberemos ser capaces de:

- Ejecutar procesos de captura de datos.
- Realizar tareas de exploración de datos.
- Gestionar la calidad de los datos, identificando problemas y proporcionando soluciones.

Preparación de datos

El objetivo final de esta fase es obtener los datos finales sobre los que aplicarán los modelos.

En esta fase deberemos ser capaces de:

- Establecer el universo de datos con los que trabajar.
- Realizar tareas de limpieza de datos.
- Construir un juego de datos apto para ser usado en modelos de minería de datos.
- Integrar datos de fuentes heterogéneas si es necesario.

Modelado

El objetivo último de esta fase es construir un modelo que nos permita alcanzar los objetivos del proyecto.

En esta fase deberemos ser capaces de:

- Seleccionar las técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos.
- Fijar una estrategia de verificación de la calidad del modelo.
- Construir un modelo a partir de la aplicación de las técnicas seleccionadas sobre el juego de datos.

- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos.

Evaluación del modelo

En esta fase nos centrarnos en evaluar el grado de acercamiento del modelo a los objetivos de negocio.

En esta fase deberemos ser capaces de:

- Evaluar el modelo o modelos generados hasta el momento.
- Revisar todo el proceso de minería de datos que nos ha llevado hasta este punto.
- Establecer los siguientes pasos a tomar, tanto si se trata de repetir fases anteriores como si se trata de abrir nuevas líneas de investigación.

(Rueda, 2019)

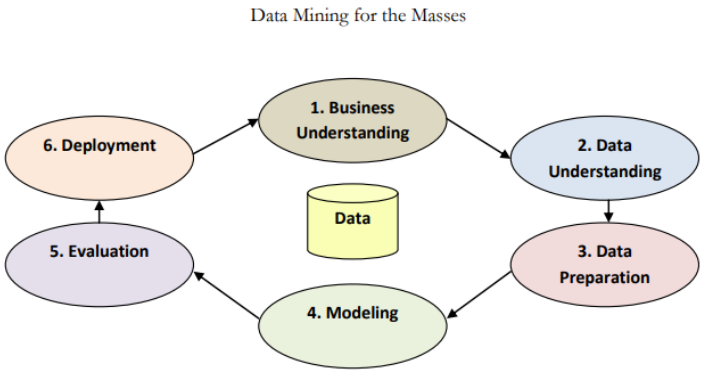


Figure 1-1: CRISP-DM Conceptual Model.

(*Data Mining for the Masses*, 2012b)

POBLACIÓN Y MUESTRA

Población:

La población estuvo constituida por el jefe de control de calidad, supervisor general.

Muestra:

Para esta investigación se tomará el 100% de la población.

Para que la muestra sea representativa se han tomado los datos de diferentes producciones dentro de la misma calidad, el tipo de muestreo será conglomerado

POBLACIÓN DE …	N
Proceso de Tostión	250
Proceso de extracción	3472
Proceso de centrifuga	4825
Proceso de Evaporación	5800
Proceso de Spray	2592
TOTAL	16939

EL TAMAÑO DE LA MUESTRA

$$n = \frac{m}{e^2 (m-1) + 1}$$

$$n = \frac{2}{(0.06)^2 (2-1) + 1}$$

$$n = \frac{2}{(0.0036)(1) + 1}$$

$$n = \frac{2}{1.0036}$$

$$n = 1.992852583$$

INSTRUMENTOS DE RECOLECCIÓN DE DATOS

El instrumento usado para la recolección de datos utilizado es la entrevista abierta, debido a que es importante tener las explicaciones especializadas de los encargados de la producción.

Resultado y análisis de entrevista

Las entrevistas fueron realizadas a 2 personas claves dentro de la producción de café, Algunas preguntas se realizaron solo al jefe de control de calidad. Para mayor comprensión a continuación en la tabla se mostrarán los roles y los cargos de las personas que fueron entrevistadas

Rol del entrevistado	Entrevista
Jefe de control de calidad	E1
Supervisor general.	E2

Tal y como se puede ver en la tabla los 2 roles de las personas a las que se les realizo la entrevista, las cuales cumplen un rol muy importante en la producción de café de la compañía ASKELGADO S.A.

Preguntas realizadas en la entrevista

¿Cómo define usted la selección de los recursos de la producción A/R?
¿Cuándo tienen una producción que parámetros toman en cuenta en cada área?
¿Con que frecuencia se realiza el análisis para la asignación de recursos?
Actualmente, ¿existe un método tecnológico para poder resolver de manera más rápida la asignación de recursos?

Están de acuerdo con la implementación de un modelo predictivo como un árbol de decisión para el análisis de los datos?

Como ya fue mencionado anteriormente existieron preguntas en común en la entrevista, las cuales nos permiten decir que los 2 expertos tanto el jefe de control de calidad como el supervisor general, coinciden en sus respuestas.

Ambos explicaron los procesos que y los recursos que se usan para el procesamiento de café en la calidad A/R sobre todo basado en los parámetros que estos deben cumplir, ya que al ser una calidad de 50% arábica y 50% robusta su recurso más importante después del recurso humano son estos 2 tipo de café.

E1: El jefe de control de calidad con respecto a su rol dentro del trabajo de investigación, pudo mencionar la importancia de la exista un recurso tecnológico, como un algoritmo de árbol, ya que estos son usados en BPM y se les puede dar otro sentido en la producción de café, lo que le permita ocuparse de otra información diferente a la de la asignación de recursos.

Actualmente el indico que hay mucha información que solo existe en papeles, debido a que no hay una persona encargada del paso de información de los reportes que se entregan en cada producción, por el tiempo que este lleva, también considera que, aunque esta calidad se produce entre 1 a 2 veces al año con posible aumento, es importante que exista información y sobre todo la importancia de que los históricos puedan ser analizados para la toma de decisiones.

E2: El supervisor general estuvo enfocado en la facilidad de la lectura de la información y los parámetros que prácticamente se toman en cuenta como los recursos a usar para poder tomar decisiones sobre la presión de las bombas, sobre si algo puede pasar en la producción y de cómo debe interpretar los datos sin necesidad de acudir a alguna ayuda extra, lo cual le permite ahorrar mucho tiempo en búsqueda de papeles. La verdad, si serviría de mucho para asignar recursos y poder llevar un control de cómo se han venido manejando las producciones pasadas.

CAPÍTULO IV
PROPUESTA TECNOLOGICA

En este capítulo se desarrollará el modelo predictivo planteado mediante el uso de minería de datos y su metodología CRISP-DM, cumpliendo cada etapa y desarrollando cada proceso propuesto en la misma. A continuación, se detallar el desarrollo en cada etapa de la metodología en la herramienta de minería de datos elegida según sus características.

HERRAMIENTAS DE DESARROLLO

La herramienta en la que se hará la implementación de la metodología CRISP-DM es en Jupyter, es una página interactiva de desarrollo, es flexible y permite desarrollar en diferentes lenguajes como Python3, C++, R, Java de manera dinámica, a la vez que integrar en un mismo documento tanto bloques de código como texto, gráficas o imágenes. Es un SaaS utilizado ampliamente en análisis numérico, estadística y machine learning, entre otros campos de la informática y las matemáticas. (*Project Jupyter*, 2020)

- Jupyter.
- Python.
- Excel.

El lenguaje de programación usado para implementar el algoritmo de minería de datos es Python, de acuerdo al problema planteado ya que en comparación a los diferentes métodos como lo son R, RapidMiner y SQL, Python permite que el aprendizaje sea más rápido y efectivo mientras que R tiene una curva de aprendizaje más lenta y complicada. También, aunque ambos lenguajes permiten realizar las mismas operaciones Python se vuelve una opción más viable por su rapidez y fácil adaptación.

(admin, 2015)

Modelo para la asignación de recursos de la calidad A/R de la compañía ASKELGADO S.A.

Se presenta un modelo para la asignación de recursos de la calidad A/R, el cual se encuentra dividido en la comprensión del negocio, el análisis y proceso de datos, las técnicas de minería, filtro y análisis de datos y la generación.

Conceptualización del modelo propuesto.

Partiendo de los términos mencionado previamente, se considera “La minería de datos para la asignación” como: el conjunto de técnicas y modelos, aplicados con el fin de optimizar los recursos y minimizar los tiempos de asignación de los mismos.

Es importante conocer que no existen soluciones únicas para la asignación de recursos para la calidad de café A/R, esto se da porque no hay un único escenario y sus variables pueden variar de acuerdo a factores internos o externos

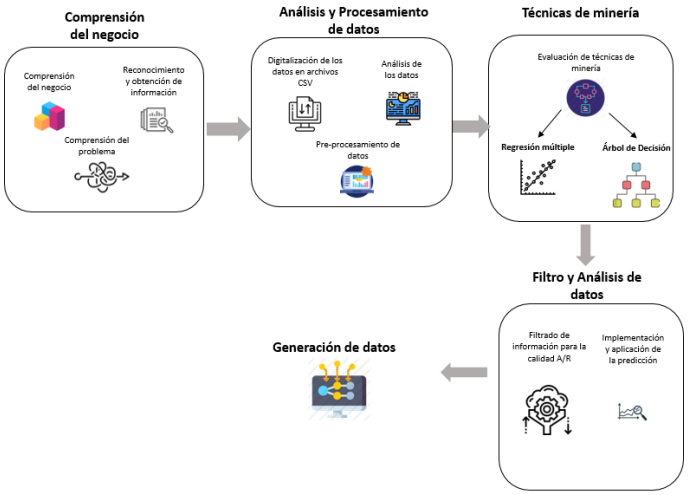


Figura 1. Representación gráfica del modelo para la asignación de recursos.

Los componentes del modelo son:

- Comprensión del negocio: Contiene la comprensión del negocio, comprensión del problema y el reconocimiento de la información. Este componente genera como salida, la obtención de los datos que se encontraban almacenados documentos físicos.

- Análisis y Procesamiento de los datos: Contiene la digitalización de los datos, el pre-procesamiento de datos y el análisis de datos. Recibe como entrada la información de las áreas del procesamiento del café y genera como salida el análisis de los datos.
- Técnicas de Minería: Contiene la evaluación de las técnicas de minería como la regresión múltiple y el árbol de decisiones. Recibe como entrada los datos que fueron analizados previamente y genera como salida la técnica que tiene mejor precisión y una regresión de la técnica para realizar la predicción.
- Filtro y Análisis de datos: Contiene la implementación y aplicación de la predicción de los datos y el filtrado de la información para la calidad A/R. Recibe como entrada la data previamente analizada y genera como salida los datos filtrados.
- Generación de datos: Recibe los datos filtrados para generar la data que será mostrada para la toma de decisiones.

Para la instrumentación del modelo se siguen los siguientes procesos

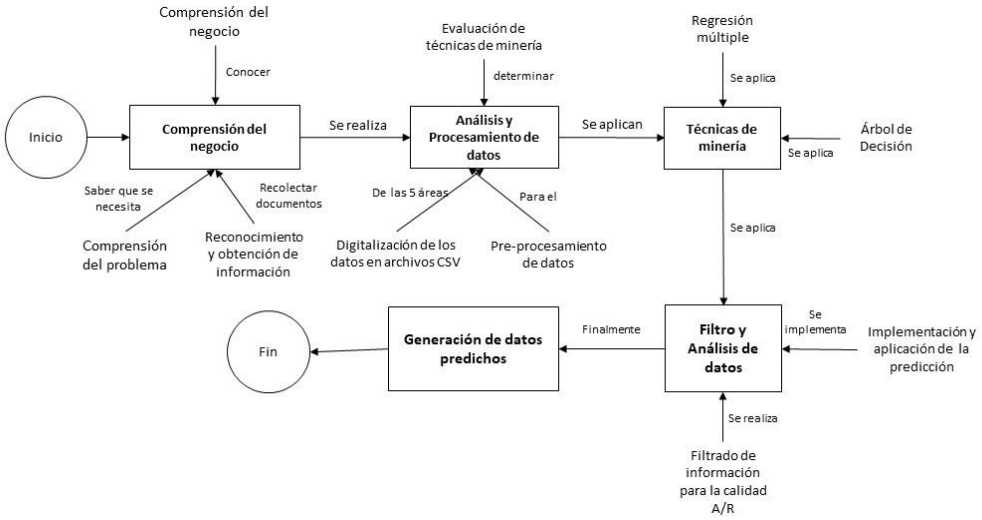


Figura 1. Instrumentación del modelo para la asignación de recursos de la calidad A/R.

Descripción de los procesos de la instrumentación del modelo.

- Comprensión del negocio: proceso que contiene como objetivo la comprensión del negocio para poder comprender el problema que se está planteando y así poder realizar un reconocimiento y obtención de información.
- Análisis y procesamiento de datos: el proceso tiene como objetivo la digitalización de los datos en archivos CSV que fueron obtenidos físicamente para así poder ser pre-procesados, también se analiza que técnica de minería se van a usar.
- Técnica de Minería: el proceso tiene como objetivo determinar cuál de las técnicas evaluadas en el proceso anterior tiene mejor rendimiento, de las cuales se tiene la regresión múltiple y el árbol de decisión.
- Filtro y análisis de datos: el objetivo de este proceso es implementar la aplicación de la predicción y filtrar la información que contendrá la calidad A/R.
- Generación de los datos predichos: el objetivo de este proceso es que una vez filtrada toda la información para calidad de café A/R se genera solo la data necesaria para su interpretación.

El procesamiento de la información de las áreas de producción permite identificar los recursos necesarios para que sean referente en la siguiente toma de decisiones, aprovechando la ganancia de tiempo en la decisión y dando pauta a mejorar.

Objetivo

Determinar los recursos que se van a utilizar para la producción de la calidad A/R, con el fin de optimizar el proceso de asignación de recursos de manera que reduzcan tiempos en la toma de decisión.

Descripción de la solución:

El programa presentara los resultados de la aplicación de las técnicas de minería de datos que determinan cuales son los recursos que se van a usar y cuáles son los valores que estos contienen para que puedan cumplir con la

calidad antes mencionada, se alimentara de varios archivos CSV que contienen las variables correspondientes a cada área y valores que estos deben cumplir.

Herramienta de Minería de Datos:

La herramienta para la programación de la minería de datos es Jupyter, la misma que se alimenta de archivos en formato CSV al cual se le aplican las técnicas escogidas, como el árbol de regresión, para determinar los recursos a usar.

Proceso

En los siguientes párrafos se presentará el proceso de las fases de la minería de datos para la asignación de recursos.

Recopilación de datos

La primera fase consiste en la recolección de la información de cada una de las áreas de producción conformadas por Tostión, Extracción, Centrifuga, Evaporación y Spray, correspondiente a los años 2018-2019 se obtuvieron de documentos físicos e informes por el jefe de producción, las cuales fueron procesadas a ser digitalizadas para en archivos de Excel con extensión CSV para poder ser procesadas.

De los cuales exiten en totalidad 53939 datos los cuales están divididos en

Cantidad total de datos		Cantidad de Variables por año		Suma de años	Cantidad total de variables por año		Cantidad total de variables
Variables	Campos	2018	2019		2018	2019	
Tostión	10	606	583	1189	6060	5830	11890
Extraction	41	242	225	467	9922	9225	19147
Centrifuga	25	142	93	235	3550	2325	5875
Evaporacion	35	200	143	343	7000	5005	12005
Spray	27	96	90	186	2592	2430	5022
	276	1286	1134	2420	29124	24815	53939

TOSTION	2018	2019	CAMPOS 2018	CAMPOS 2019
int	1	1	606	583
object	3	3	1818	1749
float	6	6	3636	3498
			6060	5830

EXTRACCION	2018	2019	CAMPOS 2018	CAMPOS 2019
int	2	2	484	450
object	3	3	726	675
float	36	36	8712	8100
			9922	9225

CENTRIFUGA	2018	2019	CAMPOS 2018	CAMPOS 2019
int	1	1	142	93
object	0	0	0	0
float	24	24	3408	2232
			3550	2325

EVAPORACION	2018	2019	CAMPOS 2018	CAMPOS 2019
int	1	1	200	143
object	0	0	0	0
float	34	34	6800	4862
			7000	5005

SPRAY	2018	2019	CAMPOS 2018	CAMPOS 2019
int	1	1	96	90
object	0	0	0	0
float	26	26	15756	15158
			15852	15248

Estos datos fueron solicitados a el Jefe de Producción y Supervisor de Planta, quienes facilitaron la información necesaria para su análisis.

Los datos obtenidos de cada una de las áreas son los siguientes:

– **Tostión**

Campos	Tipo de Datos	Descripción
batch	int64	Cuenta el número de cargas en los tostadores
tipo	object	Tipo de café que carga
hora_inicio	object	Hora de inicio de carga
hora_fin	object	Hora de finalización de carga
temp_inicial	float64	Temperatura inicial y final del tostador
agua	float64	Agua sobre segundo
k_cafe_verde	float64	Kilos de café Verde
k_cafe_tostado	float64	Kilos de café Tostado
color_pvolt	float64	Color del café tostado
h2o	float64	Porcentaje de agua

– **Extracción**

Campos	Tipo de Datos	Descripción
N_ext	int64	Numero de extracciones
extractor#	int64	Numero de extractor
carga_kg	int64	Cantidad de carga en KG
hora_arranque	object	Hora de arranque de cada extractor

h_env_balanza	object	Envió del café tostado a las balanzas
h_corte	object	Hora de corte del extractor
hidrolisis	int64	Cantidad de Hidrolisis
t_parada_min	int64	Tiempo de parada del extractor
flujo	int64	Flujo de alimentación de agua
temp_cal	int64	Temperatura del agua en el calentador
p_calentador	int64	Presión de agua en el calentador
brix	float64	Brix del cafe
solido_soluble	float64	Rendimiento del solido
rendimiento	int64	Porcentaje de rendimiento
kg._solido_soluble	int64	KG de solido soluble
ph_agua_blanda	int64	Ph. Del agua blanda
brix_bagazo	int64	Brix del bagazo
temp_ext1	int64	Temperatura del extractor 1
presion_ext1	int64	Presion del extractor 1
reflujo_ext1	int64	Reflujo del extractor 1
temp_ext2	int64	Tempreatura del extractor 2
presion_ext2	int64	Presion del extractor 2
reflujo_ext2	int64	Reflujo del extractor 2
temp_ext3	int64	Tempreatura del extractor 3

presion_ext3	int64	Presion del extractor 3
reflujo_ext3	int64	Reflujo del extractor 3
temp_ext4	int64	Tempreatura del extractor 4
presion_ext4	int64	Presion del extractor 4
reflujo_ext4	int64	Reflujo del extractor 4
temp_ext5	int64	Tempreatura del extractor 5
presion_ext5	int64	Presion del extractor 5
reflujo_ext5	int64	Reflujo del extractor 5
temp_ext6	int64	Tempreatura del extractor 6
presion_ext6	int64	Presion del extractor 6
reflujo_ext6	int64	Reflujo del extractor 6
temp_ext7	int64	Tempreatura del extractor 7
presion_ext7	int64	Presion del extractor 7
reflujo_ext7	int64	Reflujo del extractor 7
temp_ext8	int64	Tempreatura del extractor 8
presion_ext8	int64	Presion del extractor 8
reflujo_ext8	int64	Reflujo del extractor 8

- **Centrifuga**

Campos	Tipo de Datos	Descripción
horas	int64	Hora de clarificado

stock_bruto	float64	Stock de extracto bruto
centrifuga	float64	Stock de extracto de centrifuga # 1 y 2
ext_brix	float64	Brix del extracto
ext_temp	float64	Temperatura del extracto
ext_presion	float64	Presion del extracto
ext_presion.1	float64	Contrapresión de la extracion
ext_filtro	float64	Filtro del extracto
cent_rpm	float64	Centrifuga RPM
cent_aceite	float64	Aceite de la centrifuga
ctf_amp	float64	Centrifuga AMP
lavado_agua	float64	Cantidad de agua en el lavado de centrigua
lavado_quimico	float64	Quimicos del lavado
t_d_parc	float64	Tiempo de parada
t_d_total	float64	Total de tiempo
tn1	float64	Lavado de tanque 1
tn2	float64	Lavado de tanque 2
tn3	float64	Lavado de tanque 3
tn4	float64	Lavado de tanque 4
tn5	float64	Lavado de tanque 5

tn6	float64	Lavado de tanque 6
tn7	float64	Lavado de tanque 7
tn8	float64	Lavado de tanque 8
tn9	float64	Lavado de tanque 9
tn10	float64	Lavado de tanque 10

– **Evaporación**

Campos	**Tipo de Datos**	**Descripción**
Hora	object	Hora de evaporación
Tanque_N	float64	Numero de tanque de alimentación
Cantidad_lts	float64	Cantidad de litros en el tanque de alimentación
temp	float64	Temperatura del tanque de alimentación
brix	float64	Brix del tanque de alimentación de evaporación
filtro	object	Filtro del tanque de alimentación
agua_entrada_PSI	float64	Agua de la torre, entrada PSI
agua_entrada_C	float64	Agua de la torre, entrada en grados centígrados

agua_salida_PSI	float64	Agua de la torre, salida PSI
agua_salida_C	float64	Agua de la torre, salida, en grados centígrados
alim_extr_al_evapor/hora	float64	Alimentación de flujo del extracto al evaporador para la evaporación por hora
bomba_frec_entrada	float64	Bomba de frecuencia de entrada de la alimentación del flujo
bomba_frec_salida	float64	Bomba de frecuencia de salida de la alimentación del flujo
Vacio_HG	float64	Bar de entrada de alimentación del vapor
bomba_vacio_IN	float64	Vacío IN de alimentación del vapor
calent_C	float64	Temperatura del calentador en grados centígrados
ef1_calandrina	float64	Grados centígrados de calandria del efecto 1
ef1_producto	float64	Grados centígrados del producto del efecto 1
ef1_vacio_IN	float64	Vacío IN del efecto 1
ef1_filtro	float64	Filtro del efecto 1

ef1_brix	float64	Brix del efecto 1
ef2_calandrina	float64	Grados centígrados de calandria del efecto 2
ef2_producto	float64	Grados centígrados del producto del efecto 2
ef2_vacio_IN	float64	Vacio IN del efecto 2
ef2_filtro	float64	Filtro del efecto 2
ef2_brix	float64	Brix del efecto 2
b-1	float64	Sello de las bombas de agua, bomba 1
b-2	float64	Sello de las bombas de agua, bomba 2
b-3	float64	Sello de las bombas de agua, bomba 3
b-4	float64	Sello de las bombas de agua, bomba 4
b-5	float64	Sello de las bombas de agua, bomba 5
ext_hora	float64	Hora de extracto concentrado
ext_tanq	float64	Numero de tanque de extracto de concentrado
ext_concentrado	float64	Concentrado del extracto

ext_brix	float64	Brix del extracto concentrado
ext_filtro	float64	Filtro del extracto del concentrado

– **Spray**

Campos	**Tipo de Datos**	**Descripción**
hora	Int64	Hora de spray
Stock_Extracto_spray	float64	Stock de extracto en spray
ext_brix	float64	Brix del extracto
ext_temp	float64	Temp extracto
temp_placa	float64	Temperatura de la placa extracto
presion_bomba	float64	Presión de las toberas del extracto
presion_toberas	float64	Temperatura de entrada de aire
aire_temp_entrada	float64	Temperatura de salida de aire
aire_temp_salida	float64	Temperatura de aire del cono
aire_temp_cono	float64	Abertura del ventilador del aire
aire_abertura_vent	float64	Abertura de damper del aire

aire_abertura_damp	float64	Vacio del aire
aire_vacio	float64	Retorno de finos del aire
aire_retorno_finos	float64	Retorno de finos aire
aire_humos_comb	float64	Humos y combustible de aire
horno_tobera_uso	float64	Tobera en uso horno
hora_presion_entra	float64	Presion entrada del horno
horno_presion_retorno	float64	Presion de retorno horno
horno_volumen_tanque	float64	Volumen tanque del horno
inyeccion_co2	float64	Inyeccion de Co2
equipos_bombas	float64	Bombas
equipos_motores	float64	Motores
equipos_toberas_uso	float64	Toberas en uso
equipos_tableros	float64	tableros
soluble	float64	Soluble
soluble_total	float64	Total de soluble

Pre-Procesamiento de datos.

La fase de pre-procesamiento permite determinar cuáles son los datos que mayor relevancia tienen.

Se procedieron a llamar los datos de Tostión, extracción, centrifuga, evaporación y spray de los años 2018-2019, los cuales fueron ingresados en

el programa Fig 1 Clase cargardatos, estos datos fueron ingresados con datos de columnas que el experto considero como primera instancia los más importantes como un primer filtro. De los cuales cada área se quedó con los siguientes datos:

Tostión: café verde, café tostado y el color del café, Fig 2 datos de la clase cargardatos Tostión, Fig 3 datos de la clase cargardatos tostión.

Extracción: hidrolisis, brix, temperatura y presión, Fig 4 datos de la clase cargardatos extracción, Fig 5 datos de la clase cargardatos extracción.

Centrifuga: al stock bruto y temperatura, Fig 6 datos de la clase cargardatos centrifuga, Fig 7 datos de la clase cargardatos centrifuga.

Evaporación: cantidad en litros, el efecto, el efecto del brix y el brix de la extracción, Fig 8 datos de la clase cargardatos evaporación, Fig 9 datos de la clase cargardatos evaporación.

Spray: stock del extracto y el brix, Fig 10 datos de la clase cargardatos spray, Fig 11 datos de la clase cargardatos spray.

Los mismos a los cuales se les realizo una filtración para verificar que los datos que se encuentran dentro de esas variables sean de tipo flotante para poder crear un nuevo dataframe que contendrá el nombre del área con los datos pre-procesados. De los cuales se unieron y se obtuvo una tabla con datos que se encontraban con valores nulos por la diferencia en las columnas, Fig 12 datos de la clase cargardatos.

Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_tostado	k_cafe_verde	presion_ext1	stoc
NaN	NaN	NaN	15.0	NaN	NaN	NaN	NaN	NaN	67.35	185.6	NaN	
NaN	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN	67.35	185.6	NaN	
NaN	NaN	NaN	18.0	NaN	NaN	NaN	NaN	NaN	67.35	185.6	NaN	
NaN	NaN	NaN	19.0	NaN	NaN	NaN	NaN	NaN	67.35	185.6	NaN	
NaN	NaN	NaN	21.0	NaN	NaN	NaN	NaN	NaN	67.35	185.6	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	
0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	
0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	
0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	
9000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	NaN	NaN	NaN	
7000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	NaN	NaN	NaN	

ws × 14 columns

Los datos nulos o NaN del dataframe fueron sustituidos por los valores de la media, Fig 14 datos de la clase datosmedia.

	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_tostado	k_cafe_verde	presion_e
0	3200.0	14.2	1500.0	15.0	18.0	82.0	50.0	50.0	2086.0	67.35	185.6	1
1	3200.0	14.2	1500.0	20.0	18.0	82.0	50.0	50.0	2086.0	67.35	185.6	1
2	3200.0	14.2	1500.0	18.0	18.0	82.0	50.0	50.0	2086.0	67.35	185.6	1
3	3200.0	14.2	1500.0	19.0	18.0	82.0	50.0	50.0	2086.0	67.35	185.6	1
4	3200.0	14.2	1500.0	21.0	18.0	82.0	50.0	50.0	2086.0	67.35	185.6	1
...	...	...	...	...	...	...	...	...	...	...	...	
2547	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	101.02	287.4	1
2548	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	101.02	287.4	1
2549	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	101.02	287.4	1
2550	9000.0	14.2	1500.0	22.0	18.0	82.0	52.0	50.0	2086.0	101.02	287.4	1
2551	7000.0	14.2	1500.0	22.0	18.0	82.0	52.0	50.0	2086.0	101.02	287.4	1

2552 rows × 14 columns

Una vez generada esta información, se procedió a realizar otra reducción de variables, ya que el algoritmo contaba con muchas variables para su análisis, para esta disminución se procedió a usar 2 métodos, el primero es la matriz, Fig 14 datos de la clase datosmedia, matriz, de correlación que bajo observación nos permite ver cuál de los datos es el más aproximado contra la variable que es contrastada, en este caso el café Verde y el segundo criterio fue realizado bajo el análisis del experto. Dejando así 7 variables que van a conformar el nuevo DataFrame, Fig 16 datos de la clase datosmedia.

	k_cafe_verde	k_cafe_tostado	color_pvolt	temp_ext1	ext_temp	ef2_producto	ext_brix
0	185.6	67.35	15	142	50	82	50
1	185.6	67.35	20	142	50	82	50
2	185.6	67.35	18	142	50	82	50
3	185.6	67.35	19	142	50	82	50
4	185.6	67.35	21	142	50	82	50
...	...	...	...	...	...	...	...
2547	287.4	101.02	22	142	50	82	0
2548	287.4	101.02	22	142	50	82	0
2549	287.4	101.02	22	142	50	82	0
2550	287.4	101.02	22	142	50	82	52
2551	287.4	101.02	22	142	50	82	52

Algoritmo de minería de datos.

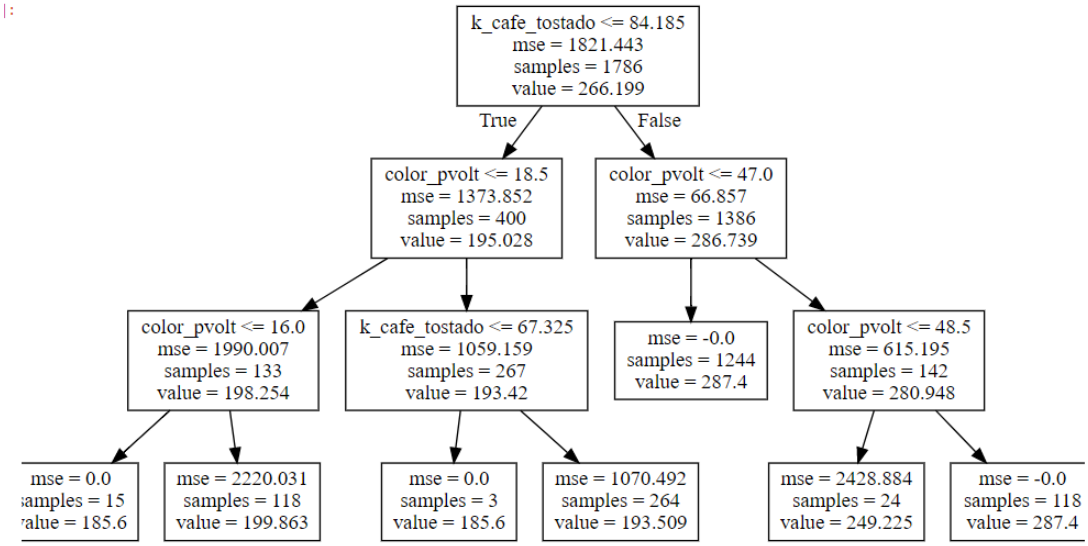
Para la comprobación y verificación de que técnica de minería se debería usar para el modelo propuesto se usaron los arboles de decisión y la regresión lineal múltiple.

ARBOL DE DECISION

Para la lectura de la información se decidió realizar un árbol de regresión el cual estará contrastado con su variable Café Verde.

Se llamaron los datos que se encontraban seleccionados de la matriz de correlación para poder realizar el estudio, Fig 17 datos de la clase árbol

El árbol fue probado con varias profundidades y una de las mejores calificaciones se dan cuando este posee 3 ramas, ya que a partir de 3 hasta 9 ramas su score se mantiene en 80%, Fig 18 datos de la clase árbol, previo a eso con 2 ramas el árbol da un score 57%, Fig 19 datos de la clase árbol



La interpretación del árbol de decisión sería: Si el valor de k_cafe_tostado <= 84.185 es verdadero predice que el valor del café verde es de 195.028, tomado de 1786 muestras que satisfacen la necesidad el primer nodo, el segundo nodo dice que si color_pvolt <= 18.5 es verdadero se predice que el valor de café verde es de 198.254 tomado de 400 muestras que satisfacen la necesidad del segundo, si es verdadero el tercer nodo dice que si el color_pvolt <= 16.0 si es verdadero se predice que el valor del café verde toma un valor de 185.6, en caso de ser falso toma el valor de 199.863, si esto se encuentra en un intervalo de 18.5 a 16.0, en caso de que el segundo nodo sea falso se predice un valor de 193.42 para el café verde tomado de 400 muestras que satisfacen las necesidades de ese nodo, si el k_cafe_tostado <= 67.325, si es verdadero se predice que el valor del café verde es 185.6, tomado de 267 muestras, en caso de ser falso su valor es de 193.509, si esto se encuentra en un intervalo de 67.325, 84.185, si el primer nodo es falso se predice que el color del café verde es de 286.739 tomado de una muestra de

1786, si el color_pvolt <= 47.0 se predice que el café verde toma un valor de 287.4, tomado de 1386 muestras, caso contrario el café verde toma el valor de 280.948, si el color_pvolt <= 48.5, si es verdadero el café verde adquiere un valor de 249.225 tomado de 142 muestras, en caso de ser falso toma el valor de 287.4, si estos se encuentran en un intervalo de 48.5 y 47.0

Regresión Múltiple

Para la regresión múltiple se realizó el entrenado y medición de los datos, el cual retorna un coeficiente de precisión del 82%, Fig 20 datos de la clase regresiónmúltiple.

La técnica que retorna una mejor precisión es el árbol de decisión, el cual se toma para poder realizar la predicción de los datos, en los cuales además de la interpretación del grafico también se predicen los valores que debe tener cada una de las áreas, Fig 24 datos de la clase modelo y devuelve los siguientes valores correspondientes a la calidad A/R:

Variable	Valores Posibles	
Café Verde	>= 185	<= 287.4
Café Tostado	>= 67.32	<= 101.02
Color del Café	>= 16	<= 48
Temperatura de Extraccion	>= 121.1	<= 142
Temperatura de Centrifuga	>= 44.62	<= 50
Efecto de Evaporacion	>= 26.59	<= 86.47
Brix de Spray	>= 46.84	<= 55.06

Interpretación del modelo predictivo:

Se toman como variables los siguientes campos:

k_cafe_tostado, color_pvolt, k_cafe_verde, temp_ext1, ext_temp, ef2_product, ext_brix.

Despues del algoritmo aplicado, estas son las variables que determinan el modelo, debido a que los factores determinantes y valores relevantes las mismas que brindan los recursos y los valores que estos van a tener.

La generación a partir de los datos procesados y que forman parte del modelo propuesto tienen como resultado 5 condiciones para poder cumplir con la calidad A/R, las cuales son correspondientes a cada area:

k_cafe_tostado > 67.32 < 101.02
color_pvolt > 17 < 48
temp_ext1 > 121.1 < 142
ext_temp > 44.62 < 50
ef2_product > 26.59 < 86.71
ext_brix > 46.84 < 55.06
Calidad A/R

Si k_cafe_tostado es > 67.32 < 101.02, color_pvolt > 17 < 48, temp_ext1 > 121.1 < 142, temp_ext1 > 121.1 < 142, ext_temp > 44.62 < 50, ef2_product > 26.59 < 86.71 y ext_brix > 46.84 < 55.06 entonces se cumplen los requerimientos para la calidad A/R.

Este es el conjunto de reglas son el resultado del uso de las técnicas de minería de datos, luego de un proceso completo, que generan un modelo predictivo para la producción de café de la calidad A/R

CONCLUSIONES

Es razonable, concluir que en el proceso de investigación se entrevistaron al Jefe de Producción y al Supervisor de planta, a través de las cuales se logró identificar los recursos que se van a usar para la producción, también se pudo reconocer la técnica de minería de datos, el modelo predictivo a aplicar y la evaluación del mismo algoritmo. La metodología identificada mediante la entrevista fue la documental para la recolección de los datos para poder implementar la minería de datos para la asignación de recursos es CRISP-MP.

Finalmente, se propone un modelo de árbol de decisión de regresión que contiene la información de las 5 áreas más importantes en la producción de café para poder realizar el proceso de minería.

RECOMENDACIONES

Se sugiere que exista más continuidad de los datos de cada área a nivel digital, ya que información ingresada es la final y no la de cada reporte, el cual hace lento el proceso de búsqueda, ya que estos deben ser ingresados de manera manual al momento de realizar el algoritmo. También se recomienda que todos los informes se almacenen en un archivo CSV en Excel.

Sería importante considerar la asignación de 1 o 2 personas que se puedan capacitar para darle una mejora constante al algoritmo o mantener actualizado los datos del mismo.

REFERENCIAS BIBLIOGRAFICAS

11.9.1-Calidad del café-Robusta—La especie. (n.d.). Retrieved December 15, 2020, from https://www.intracen.org/guia-del-cafe/calidad-del-cafe/Robusta-la-especie/

admin. (2015, May 30). R es la herramienta más popular para la minería y la ciencia de datos. *PeruStat Analytics*. https://perustat.com/blog/r-es-la-herramienta-mas-popular-para-la-mineria-y-la-ciencia-de-datos/

administrador. (2019, August 13). Toma de decisiones en una empresa. *Circulantis*. https://circulantis.com/blog/toma-decisiones-empresa/

Análisis bayesiano para un modelo de regresión logística—MATLAB & Simulink Example—MathWorks América Latina. (2019). https://la.mathworks.com/help/stats/examples/bayesian-analysis-for-a-logistic-regression-model.html

Análisis estadístico ¿Qué es? (2020). https://www.sas.com/es_cl/insights/analytics/statistical-analysis.html

Análisis predictivo: Tres cosas que es necesario saber. (n.d.). Retrieved December 19, 2020, from https://es.mathworks.com/discovery/predictive-analytics.html

Arboles de decision y Random Forest. (2018). https://bookdown.org/content/2031/arboles-de-decision-parte-i.html

Aveiro, C. (n.d.). *Qué son los algoritmos de búsqueda?* Aveiro Peroni Estudio. Retrieved January 3, 2021, from https://aveiroperoni.com/que-son-los-algoritmos-de-busqueda/

Ch01.pdf. (n.d.). Retrieved December 15, 2020, from https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf

Computación Gráfica—EcuRed. (n.d.). Retrieved January 3, 2021, from https://www.ecured.cu/Computaci%C3%B3n_Gr%C3%A1fica

Corvo, H. S. (2019, September 8). Factibilidad: Tipos, estudio, ejemplos. *Lifeder*. https://www.lifeder.com/factibilidad/

¿Cuánta información y datos generamos al año en el mundo? (n.d.). Retrieved January 3, 2021, from https://blog.orange.es/red/datos-mundo/

Data Mining for the Masses. (2012a). https://sites.google.com/site/dataminingforthemasses/

Data Mining for the Masses. (2012b). https://sites.google.com/site/dataminingforthemasses/

Duarte, Y. A. P. (2002). *CARACTERIZACIÓN FÍSICA DE CAFÉ SEMITOSTADO*. 189.

Envíos de café, con la peor cifra desde 2013 en Ecuador | El Comercio. (n.d.). Retrieved December 15, 2020, from https://www.elcomercio.com/actualidad/envios-cafe-cifra-ecuador-caida.html

Febles Rodríguez, J. P., & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. *ACIMED*, *10*(2), 69–76.

Good hygiene practices. (n.d.). Retrieved December 15, 2020, from http://www.ico.org/projects/good-hygiene-practices/cnt/cnt_sp/sec_1/c05.coffeemarkets.html

Heras, J. M. (2020). Máquinas de Vectores de Soporte (SVM). *IArtificial.net*. https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Hernández, F. (n.d.). *1 Árboles de regresión | Modelos Predictivos*. Retrieved February 17, 2021, from https://fhernanb.github.io/libro_mod_pred/arb-de-regre.html

Historia del café—Descubre todo sobre sus orígenes. (2015, September 4). *Mundo del Café*. https://mundodelcafe.es/historia-del-cafe/

Impulso de gradiente—Lo que necesitas saber—Aprendizaje automático. (2020, August 5). *DATA SCIENCE*. https://datascience.eu/es/aprendizaje-automatico/impulso-de-gradiente-lo-que-necesitas-saber/

Infante, M., Abreu, Y., Delgado, M., & Infante, O. (2019). *Minería tecnológica para el análisis de oportunidades de publicaciones en la universidad*. 16.

Inteligencia artificial – Qué es y por qué es importante. (2020). https://www.sas.com/es_cl/insights/analytics/what-is-artificial-intelligence.html

Jiawei Han, M. K. (2006a). *Ch01.pdf*. https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf

Jiawei Han, M. K. (2006b). *Data Mining _ Concepts and Techniques Solution Manual (PDFDrive).pdf*. Elsevier.

Julius T Tou y Rafael C. Gonzalez. (2019). *Reconocimiento de patrones*. http://profesores.fi-b.unam.mx/ana/APUNTES_RP/capitulo1.pdf

Kanniah, J. (2020, August 19). “Café 100% Arábica”: ¿Qué Significa? *Perfect Daily Grind Español*. https://perfectdailygrind.com/es/2020/08/19/cafe-100-arabica-que-significa/

La cosecha y los tipos de granos de café. (n.d.). Philips. Retrieved December 15, 2020, from https://www.philips.cl/c-m-ho/cafe/cafe-101/la-cosecha-y-los-tipos-de-granos-de-cafe

La historia del café | Nescafe Argentina. (2015). https://www.nescafe.com/ar/la-historia-del-cafe

Latimore, Z. (2019, October 16). Cómo Usar Los Datos de Tueste Del Café: RoR, Temperatura y Más. *Perfect Daily Grind Español*. https://perfectdailygrind.com/es/2019/10/17/como-usar-los-datos-de-tueste-del-cafe-ror-temperatura-y-mas/

Los métodos del Data Mining o Minería de datos. (n.d.). *Los Métodos Del Data Mining o Minería de Datos*. Retrieved December 21, 2020, from http://traduccionesbigdata.blogspot.com/2017/07/los-metodos-del-data-mining-o-mineria.html

Maria Consuelo Justicia de la Torre. (2017). *NUEVAS TÉCNICAS DE MINERÍA DE TEXTOS: APLICACIONES*. 297.

Método Cualitativo—Concepto, características y ejemplos. (2019). https://concepto.de/metodo-cualitativo/

Metodología de la investigación (definición y conceptos). (2018). *Web y Empresas*. https://www.webyempresas.com/metodologia-de-la-investigacion/

Minería de datos en bibliotecas: Bibliominería. (2006, December). [Text.Article]. http://bid.ub.edu/17canda2.htm

Modelado predictivo: La única guía que necesita. (n.d.). MicroStrategy. Retrieved December 15, 2020, from https://www3.microstrategy.com/es/resources/introductory-guides/predictive-modeling-the-only-guide-you-need

Morales, A. (2019, July 2). Lenguajes de programación para realizar ciencia de datos. *MappingGIS*. https://mappinggis.com/2019/07/lenguajes-de-programacion-para-realizar-ciencia-de-datos/

Posada, S. G. (2019, March 4). ¿Cómo se determina la calidad del café? *Qué Café!* https://quecafe.info/como-se-determina-la-calidad-del-cafe/

Project Jupyter. (2020). https://www.jupyter.org

Python o R. ¿Qué lenguaje utilizar para el análisis de datos? (2016, November 16). Canal Informática y TICS.

https://revistadigital.inesem.es/informatica-y-tics/python-r-analisis-datos/

¿Qué es el aprendizaje automático? | *Oracle Chile*. (n.d.). Retrieved December 19, 2020, from https://www.oracle.com/cl/data-science/machine-learning/what-is-machine-learning/

¿Qué es el método cualitativo? (2015). *Tendenzias.com*. https://tendenzias.com/ciencia/que-es-el-metodo-cualitativo/

¿Qué es el Procesamiento de Datos? | *Next U*. (2017, August 23). NextU LATAM. https://www.nextu.com/blog/que-es-el-procesamiento-de-datos/

¿Qué es una base de datos? (2020). https://www.oracle.com/mx/database/what-is-database/

Qué es y cómo interpretar una regresión logística—Incluye Ejemplo. (2019, January 23). *Conceptos Claros*. https://conceptosclaros.com/que-es-regresion-logistica/

Regla de los K vecinos más cercanos—EcuRed. (n.d.). Retrieved December 21, 2020, from https://www.ecured.cu/Regla_de_los_K_vecinos_m%C3%A1s_cercanos

Roşca, D. G., & Rădoiu, D. (2015). *STEP-BY-STEP MODEL FOR THE STUDY OF THE APRIORI ALGORITHM FOR PREDICTIVE ANALYSIS*. 4.

Rueda, J. F. V. (2019, November 4). CRISP-DM: Una metodología para minería de datos en salud. *healthdataminer.com*. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

¿Sabes en qué se diferencian las redes neuronales del Deep Learning? (2019). https://blogthinkbig.com/redes-neuronales-deep-learning

Sabora, C. (2016, June 8). *7 diferencias entre el café arábica y el café robusta.* [Text]. Cafés Sabora. https://cafesabora.com/es/7-diferencias-entre-el-caf%C3%A9-ar%C3%A1bica-y-el-caf%C3%A9-robusta

Schab, E., Rivera, R., Bracco, L., Coto, F., Cristaldo, P., Ramos, L., Rapesta, N., Núñez, J. P., Retamar, S., Casanova, C., Battista, A. D., & Herrera, N. E. (2018). *Minería de Datos y Visualización de Información*. 5.

Significado de Calidad. (n.d.). Significados. Retrieved December 15, 2020, from https://www.significados.com/calidad/

Significado de Metodología de la investigación. (2015). Significados. https://www.significados.com/metodologia-de-la-investigacion/

Spain, por K. (2020a, May 5). *Las 11 técnicas más utilizadas en el modelado de análisis predictivos*. Keyrus Spain Blog. https://keyrusspainblog.com/2020/05/05/las-11-tecnicas-mas-utilizadas-en-el-modelado-de-analisis-predictivos/

Spain, por K. (2020b, May 5). *Las 11 técnicas más utilizadas en el modelado de análisis predictivos*. Keyrus Spain Blog.

https://keyrusspainblog.com/2020/05/05/las-11-tecnicas-mas-utilizadas-en-el-modelado-de-analisis-predictivos/

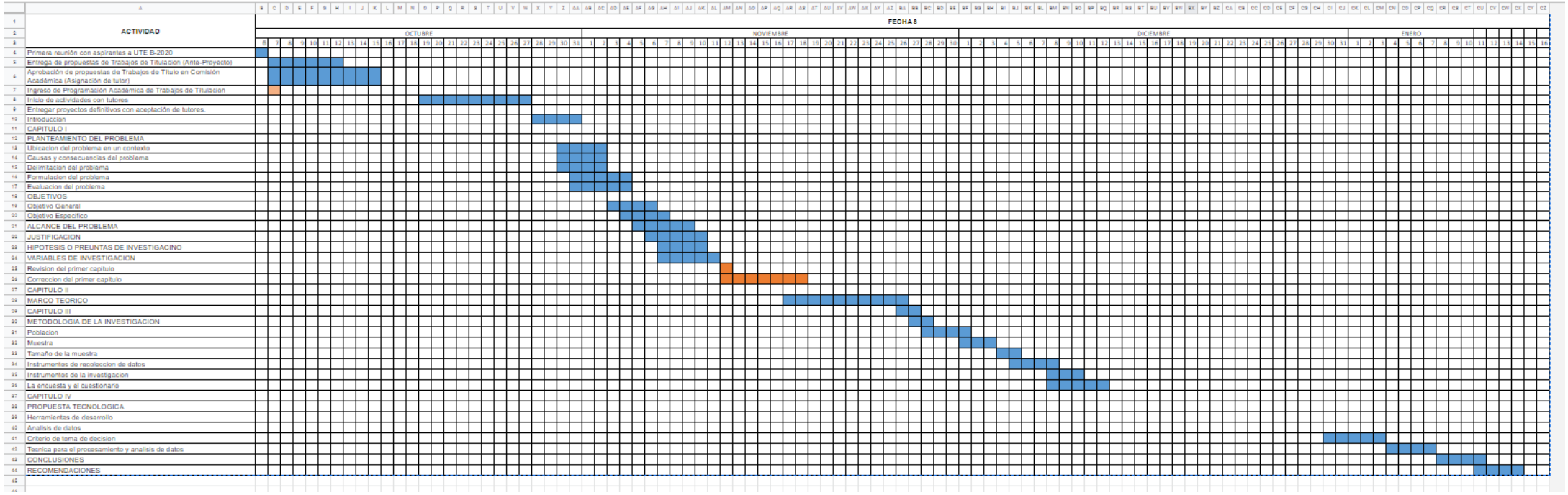
Telégrafo, E. (2017, September 15). *Las exportaciones de café bajan 17% en siete meses*. El Telégrafo. https://www.eltelegrafo.com.ec/noticias/economia/4/las-exportaciones-de-cafe-bajan-17-en-siete-meses

Timaran-Pereira, R., Calderón-Romero, A., & Hidalgo-Troya, A. (2017). Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. *Universidad y Salud*, *19*(3), 388. https://doi.org/10.22267/rus.171903.101

Tyagi, G. (2020, August 17). *Ensemble models for Classification*. Medium. https://towardsdatascience.com/ensemble-models-for-classification-d443ebed7efe

Yolanda Belinchón Monjas. (2019). *MINERÍA DE DATOS*. http://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

ANEXOS


FECHA 8
ACTIVIDAD
OCTUBRE
NOVIEMBRE
DICIEMBRE
ENERO
Primera reunión con aspirantes a UTE B-2020
Entrega de propuestas de Trabajos de Titulacion (Ante-Proyecto)
Aprobación de propuestas de Trabajos de Titulo en Comisión Académica (Asignación de tutor)
Ingreso de Programación Académica de Trabajos de Titulacion
Inicio de actividades con tutores
Entregar proyectos definitivos con aceptación de tutores.
Introduccion
CAPITULO I
PLANTEAMIENTO DEL PROBLEMA
Ubicacion del problema en un contexto
Causas y consecuencias del problema
Delimitacion del problema
Formulacion del problema
Evaluacion del problema
OBJETIVOS
Objetivo General
Objetivo Especifico
ALCANCE DEL PROBLEMA
JUSTIFICACION
HIPOTESIS O PREUNTAS DE INVESTIGACINO
VARIABLES DE INVESTIGACION
Revision del primer capitulo
Correccion del primer capitulo
CAPITULO II
MARCO TEORICO
CAPITULO III
METODOLOGIA DE LA INVESTIGACION
Poblacion
Muestra
Tamaño de la muestra
Instrumentos de recoleccion de datos
Instrumentos de la investigacion
La encuesta y el cuestionario
CAPITULO IV
PROPUESTA TECNOLOGICA
Herramientas de desarrollo
Analisis de datos
Criterio de toma de decision
Tecnica para el procesamiento y analisis de datos
CONCLUSIONES
RECOMENDACIONES

Codigo

Importar librerias

In [1]:
```
import warnings
warnings.simplefilter("ignore")
%matplotlib inline
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import graphviz
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn import preprocessing
```

Importarcion de datos

Se importan todos los datos de cada area en variables

In [2]:
```
#año 2018
tostion2018 = pd.read_csv('base/2018/tostion2018.csv', parse_dates=True, infe
extraccion2018 = pd.read_csv('base/2018/extraccion2018.csv')
centrifuga2018 = pd.read_csv('base/2018/centrifuga2018.csv')
evaporacion2018 = pd.read_csv('base/2018/evaporacion2018.csv')
spray2018 = pd.read_csv('base/2018/spray2018.csv')

#año 2019
tostion2019 = pd.read_csv('base/2019/tostion2019.csv', parse_dates=True, infe
extraccion2019 = pd.read_csv('base/2019/extraccion2019.csv')
centrifuga2019 = pd.read_csv('base/2019/centrifuga2019.csv')
evaporacion2019 = pd.read_csv('base/2019/evaporacion2019.csv')
spray2019 = pd.read_csv('base/2019/spray2019.csv')
```

Fig 1 Clase cargardatos

Tostion

In [49]:
```
# tostion año 2018 - 2019
tostion2019 = tostion2019[['k_cafe_verde','k_cafe_tostado','color_pvolt']]
tostion2018 = tostion2018[['k_cafe_verde','k_cafe_tostado','color_pvolt']]
tostion =  [tostion2019,tostion2018]
tostion = pd.concat(tostion, sort='False', ignore_index='True')
tostion
```

Out[49]:

	color_pvolt	k_cafe_tostado	k_cafe_verde
0	15.0	67.35	185.6
1	20.0	67.35	185.6
2	18.0	67.35	185.6
3	19.0	67.35	185.6
4	21.0	67.35	185.6
...	...	...	...
1183	25.0	101.02	287.4
1184	24.0	101.02	287.4
1185	21.0	101.02	287.4
1186	19.0	101.02	287.4
1187	20.0	101.02	287.4

1188 rows × 3 columns

Fig 2 datos de la clase cargardatos Tostión

In [50]:
```
#limpiamos datos de tostion
numtos = (tostion.dtypes == float) | (tostion.dtypes == int)
numtos
tostion.dtypes == object
for el in numtos.index:
    print(el)

    #extraigo solo las columnas numericas
numtos_cols = [c for c in numtos.index if numtos[c]]
numtos_cols
```

```
color_pvolt
k_cafe_tostado
k_cafe_verde
```

Out[50]:
```
['color_pvolt', 'k_cafe_tostado', 'k_cafe_verde']
```

In [51]:
```
#extraigo las columnas de texto
tostion.dtypes == object
objtos = (tostion.dtypes == object)
objtos_cols = [c for c in objtos.index if objtos[c]]
objtos_cols
tostion_num = tostion[numtos_cols]
tostion_num.describe()
```

Out[51]:

	color_pvolt	k_cafe_tostado	k_cafe_verde
count	1188.000000	1188.000000	1188.000000
mean	26.465488	84.381035	240.121044
std	11.968492	16.843397	52.853881
min	15.000000	66.000000	185.600000
25%	19.000000	67.350000	185.600000
50%	22.000000	101.020000	287.400000
75%	26.000000	101.020000	287.400000
max	54.000000	101.020000	364.000000

Fig 3 datos de la clase cargardatos tostión

Extraccion

In [52]:
```
# extraccion año 2018 - 2019
extraccion2019.dtypes
extraccion2019 = extraccion2019[['hidrolisis','brix','temp_ext1','presion_ext1']]
extraccion2018 = extraccion2018[['hidrolisis','brix','temp_ext1','presion_ext1']]
extraccion =  [extraccion2019,extraccion2018]
extraccion = pd.concat(extraccion, sort='False', ignore_index='True')
extraccion
```

Out[52]:

	brix	hidrolisis	presion_ext1	temp_ext1
0	15.4	2176.0	14.0	167.0
1	16.6	2171.0	15.0	174.0
2	14.0	2100.0	16.0	184.0
3	14.1	2079.0	0.0	0.0
4	14.7	2080.0	8.0	95.0
...	...	...	...	...
491	14.7	2075.0	0.0	0.0
492	15.3	2086.0	11.0	135.0
493	14.8	2061.0	12.0	145.0
494	15.2	2077.0	13.0	157.0
495	14.8	2079.0	14.0	169.0

496 rows × 4 columns

Fig 4 datos de la clase cargardatos extracción

In [53]:
```
#extraccion limpiamos los datos
numext = (extraccion.dtypes == float) | (extraccion.dtypes == int)
numext
extraccion.dtypes == object
for el in numext.index:
    print(el)

#extraigo solo las columnas numericas
numext_cols = [c for c in numext.index if numext[c]]
numext_cols
```

```
brix
hidrolisis
presion_ext1
temp_ext1
```

Out[53]:
```
['brix', 'hidrolisis', 'presion_ext1', 'temp_ext1']
```

In [54]:
```
#extraigo las columnas de texto
extraccion.dtypes == object
objext = (extraccion.dtypes == object)
objext_cols = [c for c in objext.index if objext[c]]
objext_cols
extraccion_num = extraccion[numext_cols]
extraccion_num.describe()
```

Out[54]:

	brix	hidrolisis	presion_ext1	temp_ext1
count	496.000000	496.000000	496.000000	496.000000
mean	14.054839	2088.814516	8.774194	104.225806
std	1.202042	26.784397	6.354941	74.843820
min	11.100000	2059.000000	0.000000	0.000000
25%	13.500000	2079.000000	0.000000	0.000000
50%	14.200000	2086.000000	12.000000	142.000000
75%	14.800000	2089.250000	14.000000	167.000000
max	16.600000	2176.000000	16.000000	185.000000

Fig 5 datos de la clase cargardatos extracción

Centrifuga

In [55]:
```
# centrifuga año 2019
centrifuga2019 = centrifuga2019[['stock_bruto','ext_temp']]
centrifuga2018 = centrifuga2018[['stock_bruto','ext_temp']]
centrifuga =  [centrifuga2019,centrifuga2018]
centrifuga = pd.concat(centrifuga, sort='False', ignore_index='True')
centrifuga
```

Out[55]:

	ext_temp	stock_bruto
0	0.0	0.0
1	50.0	0.0
2	50.0	0.0
3	50.0	11500.0
4	50.0	18700.0
...	...	...
328	0.0	0.0
329	50.0	1000.0
330	50.0	0.0
331	50.0	0.0
332	50.0	0.0

333 rows × 2 columns

Fig 6 datos de la clase cargardatos centrifuga

In [56]:
```
#limpiamos datos de centrifuga
numcent = (centrifuga.dtypes == float) | (centrifuga.dtypes == int)
numcent
centrifuga.dtypes == object
for el in numcent.index:
    print(el)

#extraigo solo las columnas numericas
numcent_cols = [c for c in numcent.index if numcent[c]]
numcent_cols
```

```
ext_temp
stock_bruto
```

Out[56]:
```
['ext_temp', 'stock_bruto']
```

In [57]:
```
#extraigo las columnas de texto
centrifuga.dtypes == object
objcent = (centrifuga.dtypes == object)
objcent_cols = [c for c in objcent.index if objcent[c]]
objcent_cols
centrifuga_num = centrifuga[numcent_cols]
centrifuga_num.describe()
```

Out[57]:

	ext_temp	stock_bruto
count	333.000000	333.000000
mean	42.597598	6594.894895
std	17.946379	9024.382855
min	0.000000	0.000000
25%	50.000000	0.000000
50%	50.000000	600.000000
75%	50.000000	10000.000000
max	55.000000	30000.000000

Fig 7 datos de la clase cargardatos centrifuga

Evaporacion

In [58]:
```
# evaporacion año 2018 - 2019
evaporacion2019.dtypes
evaporacion2019 = evaporacion2019[['cantidad_lts','ef2_producto','ef1_brix','ext_brix']]

evaporacion2018 = evaporacion2018[['cantidad_lts','ef2_producto','ef1_brix','ext_brix']]

evaporacion =  [evaporacion2019,evaporacion2018]
evaporacion = pd.concat(evaporacion, sort='False', ignore_index='True')
evaporacion
```

Out[58]:

	cantidad_lts	ef1_brix	ef2_producto	ext_brix
0	9000.0	19.0	87.0	52.0
1	7000.0	21.0	86.0	53.0
2	4000.0	20.0	86.0	52.0
3	9000.0	19.0	87.0	53.0
4	1000.0	0.0	86.0	54.0
...	...	...	...	...
338	1500.0	0.0	89.0	54.0
339	8000.0	0.0	85.0	54.0
340	2500.0	18.0	84.0	54.0
341	10000.0	21.0	84.0	55.0
342	0.0	22.0	85.0	56.0

Fig 8 datos de la clase cargardatos evaporación

In [59]:
```
#limpiamos datos de evaporacion
numev = (evaporacion.dtypes == float) | (evaporacion.dtypes == int)
numev
evaporacion2019.dtypes == object
for el in numev.index:
    print(el)
#extraigo solo las columnas numericas
numev_cols = [c for c in numev.index if numev[c]]
numev_cols
```

```
cantidad_lts
ef1_brix
ef2_producto
ext_brix
```

Out[59]:
```
['cantidad_lts', 'ef1_brix', 'ef2_producto', 'ext_brix']
```

In [61]:
```
#extraigo las columnas de texto
evaporacion.dtypes == object
obj = (evaporacion.dtypes == object)
obj_cols = [c for c in obj.index if obj[c]]
obj_cols
evaporacion_num = evaporacion[numev_cols]
evaporacion_num.describe()
```

Out[61]:

	cantidad_lts	ef1_brix	ef2_producto	ext_brix
count	340.000000	342.000000	343.000000	343.000000
mean	3545.117647	10.909357	52.230321	32.463557
std	3945.345002	10.105338	41.163943	26.527228
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	1500.000000	18.000000	82.000000	53.000000
75%	7000.000000	20.000000	85.500000	54.000000
max	11000.000000	22.000000	89.000000	56.000000

Fig 9 datos de la clase cargardatos evaporación

Spray

In [62]:
```
# spray año 2018 - 2019
spray2019 = spray2019[['Stock_Extracto_spray','ext_brix']]
spray2018 = spray2018[['Stock_Extracto_spray','ext_brix']]
spray =  [spray2019,spray2018]
spray = pd.concat(spray, sort='False', ignore_index='True')
spray
```

Out[62]:

	Stock_Extracto_spray	ext_brix
0	3000.0	50.0
1	16000.0	50.0
2	200.0	50.0
3	9000.0	50.0
4	7300.0	50.0
...	...	...
187	0.0	0.0
188	0.0	0.0
189	0.0	0.0
190	9000.0	52.0
191	7000.0	52.0

192 rows × 2 columns

Fig 10 datos de la clase cargardatos spray

In [63]:
```
#limpiamos datos de spray
numsp = (spray.dtypes == float) | (spray.dtypes == int)
numsp
spray.dtypes == object
for el in numsp.index:
    print(el)

#extraigo solo las columnas numericas
numsp_cols = [c for c in numsp.index if numsp[c]]
numsp_cols
```

```
Stock_Extracto_spray
ext_brix
```

Out[63]:
```
['Stock_Extracto_spray', 'ext_brix']
```

In [64]:
```
#extraigo las columnas de texto
spray.dtypes == object
objsp = (spray.dtypes == object)
objsp_cols = [c for c in objsp.index if objsp[c]]
objsp_cols
spray_num = spray[numsp_cols]
spray_num
```

Out[64]:

	Stock_Extracto_spray	ext_brix
0	3000.0	50.0
1	16000.0	50.0
2	200.0	50.0
3	9000.0	50.0
4	7300.0	50.0
...	...	...
187	0.0	0.0
188	0.0	0.0
189	0.0	0.0
190	9000.0	52.0
191	7000.0	52.0

192 rows × 2 columns

Fig 11 datos de la clase cargardatos spray

Union de datos

Una vez hecho el tratamiento de la informacion se unen todas las areas

In [65]:
```
dataframes =  [tostion_num,extraccion_num,centrifuga_num,evaporacion_num,spray_num]
```

In [66]:
```
datos = pd.concat(dataframes, sort='False', ignore_index='True')
datos
```

Out[66]:

	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_
0	NaN	NaN	NaN	15.0	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	18.0	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	19.0	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	NaN	21.0	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	
2547	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2548	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2549	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2550	9000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	
2551	7000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	

2552 rows × 14 columns

Guardar archivo con datos nulos previo a su llenado

In [69]:
```
datos.to_csv('1datosn.csv')
```

Fig 12 datos de la clase cargardatos

importar librerias

In [9]:
```
import warnings
warnings.simplefilter("ignore")
%matplotlib inline
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import graphviz
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn import preprocessing
```

In [16]:
```
datos = pd.read_csv('1datosn.csv')
datos
```

Out[16]:

	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_
0	NaN	NaN	NaN	15.0	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	18.0	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	19.0	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	NaN	21.0	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	
2547	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2548	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2549	0.0	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	
2550	9000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	
2551	7000.0	NaN	NaN	NaN	NaN	NaN	52.0	NaN	NaN	

2552 rows × 14 columns

Fig 13 datos de la clase datosmedia

In [22]:
```
df=pd.DataFrame(datos)
df
```

Out[22]:

	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_
0	3200.0	14.2	1500.0	15.0	18.0	82.0	50.0	50.0	2086.0	
1	3200.0	14.2	1500.0	20.0	18.0	82.0	50.0	50.0	2086.0	
2	3200.0	14.2	1500.0	18.0	18.0	82.0	50.0	50.0	2086.0	
3	3200.0	14.2	1500.0	19.0	18.0	82.0	50.0	50.0	2086.0	
4	3200.0	14.2	1500.0	21.0	18.0	82.0	50.0	50.0	2086.0	
...	...	...	...	...	...	...	...	...	...	
2547	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	
2548	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	
2549	0.0	14.2	1500.0	22.0	18.0	82.0	0.0	50.0	2086.0	
2550	9000.0	14.2	1500.0	22.0	18.0	82.0	52.0	50.0	2086.0	
2551	7000.0	14.2	1500.0	22.0	18.0	82.0	52.0	50.0	2086.0	

2552 rows × 14 columns

In [23]:
```
df.dtypes
```

Out[23]:
```
Stock_Extracto_spray    float64
brix                    float64
cantidad_lts            float64
color_pvolt             float64
ef1_brix                float64
ef2_producto            float64
ext_brix                float64
ext_temp                float64
hidrolisis              float64
k_cafe_tostado          float64
k_cafe_verde            float64
presion_ext1            float64
stock_bruto             float64
temp_ext1               float64
dtype: object
```

Fig 14 datos de la clase datosmedia

MATRIZ DE CORRELACION

In [24]:
```
correlacion = df.corr(method="pearson")
```

In [25]:
```
correlacion
```

Out[25]:

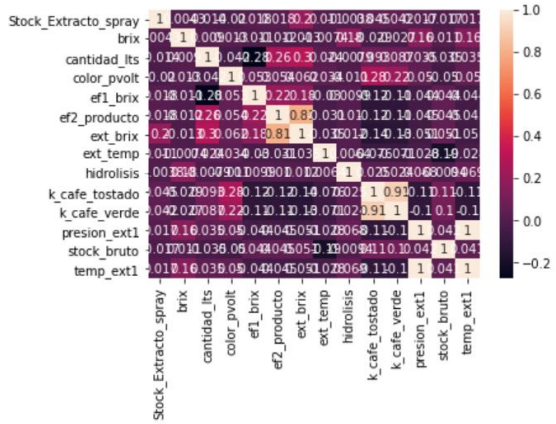
	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_
Stock_Extracto_spray	1.000000	0.004341	-0.013979	-0.020131	0.017641	0.018042	0.204303	0.01
brix	0.004341	1.000000	0.009039	0.013017	-0.011407	-0.011666	-0.013266	-0.00
cantidad_lts	-0.013979	0.009039	1.000000	-0.041915	-0.276450	0.259470	0.301804	0.02
color_pvolt	-0.020131	0.013017	-0.041915	1.000000	0.052896	0.054099	0.061518	0.03
ef1_brix	0.017641	-0.011407	-0.276450	0.052896	1.000000	0.219824	0.183518	-0.02
ef2_producto	0.018042	-0.011666	0.259470	0.054099	0.219824	1.000000	0.805614	-0.03
ext_brix	0.204303	-0.013266	0.301804	0.061518	0.183518	0.805614	1.000000	-0.03
ext_temp	0.011410	-0.007378	0.023756	0.034211	-0.029980	-0.030662	-0.034866	1.00
hidrolisis	-0.003783	0.181581	-0.007876	-0.011343	0.009940	0.010166	0.011560	0.00
k_cafe_tostado	0.044783	-0.028958	0.093245	0.278781	-0.117672	-0.120350	-0.136852	-0.07
k_cafe_verde	0.041865	-0.027071	0.087168	0.222110	-0.110004	-0.112507	-0.127934	-0.07
presion_ext1	0.016700	0.164307	0.034771	0.050074	-0.043880	-0.044879	-0.051032	-0.02
stock_bruto	-0.016733	0.010820	-0.034840	-0.050174	0.043968	0.044968	0.051134	-0.19
temp_ext1	0.016621	0.155272	0.034606	0.049837	-0.043672	-0.044666	-0.050791	-0.02

Correlacion

0.00 - 0.09 Nula
0.10 - 0.19 Muy Debil
0.20 - 0.49 Debil
0.50 - 0.69 Moderada
0.70 - 0.84 Significativa
0.85 - 0.95 Fuerte
0.96 - 1.00 Perfecta

Fig 15 datos de la clase datosmedia, matriz

In [44]:
```
df_small = df.iloc[:,:15]
correlation_mat = df_small.corr()
sns.heatmap(correlation_mat, annot = True)
plt.show()
```



In [27]:
```
#guardo mi archivo para poder visualizarlo mejor en excel para sacar la correlacion
datos.to_csv('6correlacionmedia.csv')
```

Fig 16 datos de la clase datosmedia, matriz

Los datos fueron considerados a traves de la matriz de correlacion y lo conversadocon el experto

In [31]:
```
datos = pd.read_csv('6correlacionmedia.csv')
datos = datos[['k_cafe_verde','k_cafe_tostado','color_pvolt','temp_ext1','ext_temp',
               'ef2_producto','ext_brix']]
dataframes = datos
datos
```

Out[31]:

	k_cafe_verde	k_cafe_tostado	color_pvolt	temp_ext1	ext_temp	ef2_producto	ext_brix
0	185.6	67.35	15.0	142.0	50.0	82.0	50.0
1	185.6	67.35	20.0	142.0	50.0	82.0	50.0
2	185.6	67.35	18.0	142.0	50.0	82.0	50.0
3	185.6	67.35	19.0	142.0	50.0	82.0	50.0
4	185.6	67.35	21.0	142.0	50.0	82.0	50.0
...	...	...	...	...	...	...	...
2547	287.4	101.02	22.0	142.0	50.0	82.0	0.0
2548	287.4	101.02	22.0	142.0	50.0	82.0	0.0
2549	287.4	101.02	22.0	142.0	50.0	82.0	0.0
2550	287.4	101.02	22.0	142.0	50.0	82.0	52.0
2551	287.4	101.02	22.0	142.0	50.0	82.0	52.0

2552 rows × 7 columns

In [32]:
```
#guardo mi archivo para poder visualizarlo mejor en excel para sacar la correlacion
datos.to_csv('7datosseleccionmedia.csv')
datos
```

Out[32]:

	k_cafe_verde	k_cafe_tostado	color_pvolt	temp_ext1	ext_temp	ef2_producto	ext_brix
0	185.6	67.35	15.0	142.0	50.0	82.0	50.0
1	185.6	67.35	20.0	142.0	50.0	82.0	50.0
2	185.6	67.35	18.0	142.0	50.0	82.0	50.0
3	185.6	67.35	19.0	142.0	50.0	82.0	50.0
4	185.6	67.35	21.0	142.0	50.0	82.0	50.0
...	...	...	...	...	...	...	...

Fig 17 datos de la clase datosmedia

importar librerias

In [1]:
```
import warnings
warnings.simplefilter("ignore")
%matplotlib inline
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import graphviz
## Este proporciona funciones para la estimación de muchos modelos estadísticos
import statsmodels.api as sm
## Permite ajustar modelos estadísticos utilizando fórmulas de estilo R
import statsmodels.formula.api as smf
from sklearn import preprocessing
```

Arbol de desicion

In [21]:
```
datos = pd.read_csv('6correlacionmedia.csv')
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor(max_depth=3)
from sklearn.model_selection import train_test_split
datos
```

Out[21]:

	Stock_Extracto_spray	brix	cantidad_lts	color_pvolt	ef1_brix	ef2_producto	ext_brix	ext_temp	hidrolisis	k_cafe_
0	3200	14.2	1500	15	18	82	50	50	2086	
1	3200	14.2	1500	20	18	82	50	50	2086	
2	3200	14.2	1500	18	18	82	50	50	2086	
3	3200	14.2	1500	19	18	82	50	50	2086	
4	3200	14.2	1500	21	18	82	50	50	2086	
...	...	...	...	...	...	...	...	...	...	...
2547	0	14.2	1500	22	18	82	0	50	2086	
2548	0	14.2	1500	22	18	82	0	50	2086	
2549	0	14.2	1500	22	18	82	0	50	2086	
2550	9000	14.2	1500	22	18	82	52	50	2086	
2551	7000	14.2	1500	22	18	82	52	50	2086	

2552 rows × 14 columns

Fig 18 datos de la clase árbol

In [22]:
```
col_names = datos
feature_cols = ['k_cafe_verde','k_cafe_tostado','color_pvolt','temp_ext1','ext_temp',
                'ef2_producto','ext_brix']
```

In [23]:
```
X=datos[feature_cols]
y=X['k_cafe_verde'] #target es la columna de score
X = X.drop('k_cafe_verde',axis=1) #ocurre en el eje1 #se sacan los datos que no son relevantes
```

In [24]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
```

In [25]:
```
model.fit(X_train,y_train)
```

Out[25]:
```
DecisionTreeRegressor(max_depth=3)
```

In [26]:
```
model.score(X_test,y_test)
```

Out[26]:
```
0.8777596902150654
```

In [30]:
```
import graphviz
from sklearn.tree import export_graphviz
treedot = export_graphviz(model,out_file=None, feature_names=X.columns)
treedot
graphviz.Source(treedot)
```

Fig 19 datos de la clase árbol

```
from sklearn.tree import export_graphviz
treedot = export_graphviz(model,out_file=None, feature_names=X.columns)
treedot
graphviz.Source(treedot)
```

Out[30]:

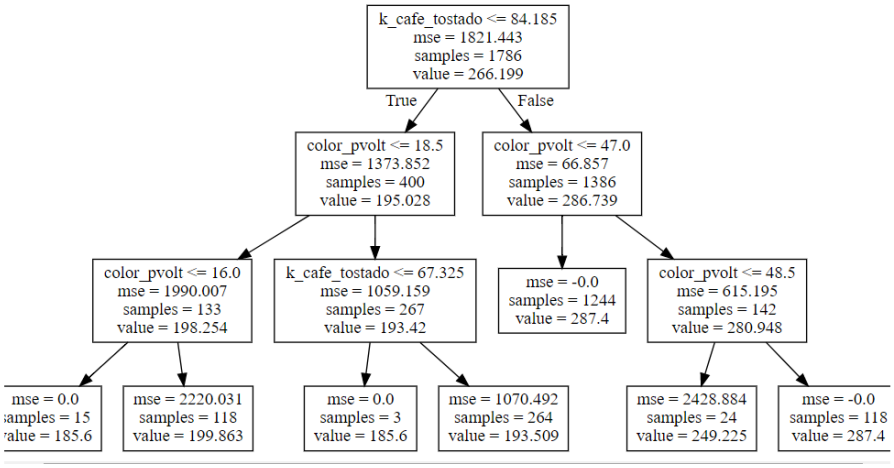


Fig 20 datos de la clase árbol

importar librerias

In [41]:
```
# Imports necesarios
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

In [42]:
```
datos = pd.read_csv('7datosseleccionmedia.csv')
dt = datos
df = dt
dt.head()
```

Out[42]:

	k_cafe_verde	k_cafe_tostado	color_pvolt	temp_ext1	ext_temp	ef2_producto	ext_brix
0	185.6	67.35	15	142	50	82	50
1	185.6	67.35	20	142	50	82	50
2	185.6	67.35	18	142	50	82	50
3	185.6	67.35	19	142	50	82	50
4	185.6	67.35	21	142	50	82	50

In [43]:
```
cafe_verde = ['k_cafe_tostado','color_pvolt','temp_ext1','ext_temp','ef2_producto','ext_brix']
X = datos[cafe_verde]
y = datos.k_cafe_verde
from sklearn.linear_model import LinearRegression
mlr_model = LinearRegression()
mlr_model.fit(X, y)
```

Out[43]:
```
LinearRegression()
```

In [44]:
```
r_sq = mlr_model.score(X, y)
print('coefficient de presicion:', r_sq)
```

```
coefficient de presicion: 0.8263138290844715
```

In [45]:
```
# Veamos los coeficienetes obtenidos, En nuestro caso, serán la Tangente
print('Coefficients: \n', mlr_model.coef_)
```

```
Coefficients:
 [ 2.78955398e+00 -1.71618121e-01 -2.76593410e-04 -1.01183432e-03
 -1.39963103e-04 -7.66158204e-04]
```

In [46]:
```
y_pred = mlr_model.predict(X)
```

In [40]:
```
y_pred
```

Out[40]:
```
array([194.62566258, 193.76757197, 194.11080821, ..., 287.38692602,
       287.34708579, 287.34708579])
```

In [39]:
```
# Este es el valor donde corta el eje Y (en X=0)
print('variable interceptora: \n', mlr_model.intercept_, )
```

```
Independent term:
 9.463126958731578 variable interceptora
```

Fig 21 datos de la clase regresiónmúltiple

importar librerias

In [39]:
```
import warnings
warnings.simplefilter("ignore")
%matplotlib inline
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import graphviz
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
```

In [40]:
```
datos = pd.read_csv('7datosseleccionmedia.csv')
df = datos
df
```

Out[40]:

	k_cafe_verde	k_cafe_tostado	color_pvolt	temp_ext1	ext_temp	ef2_producto	ext_brix
0	185.6	67.35	15	142	50	82	50
1	185.6	67.35	20	142	50	82	50
2	185.6	67.35	18	142	50	82	50
3	185.6	67.35	19	142	50	82	50
4	185.6	67.35	21	142	50	82	50
...	...	...	...	...	...	...	...
2547	287.4	101.02	22	142	50	82	0
2548	287.4	101.02	22	142	50	82	0
2549	287.4	101.02	22	142	50	82	0
2550	287.4	101.02	22	142	50	82	52
2551	287.4	101.02	22	142	50	82	52

2552 rows × 7 columns

In [16]:
```
#datos['ext_brix'].hist()
```

In [41]:
```
model = DecisionTreeRegressor(max_depth=7)
col_names = datos
feature_cols = ['k_cafe_verde','k_cafe_tostado','color_pvolt','temp_ext1','ext_temp','ef2_producto','ext_brix']
#cafe verde
X=datos[feature_cols]
y=X['k_cafe_verde']
X = X.drop('k_cafe_verde',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pverde = model.score(X_test,y_test)
av = 1 - pverde
avs = av * 100
pred_cafe_verde = model.predict(X_test)
#cafe tostado
X=datos[feature_cols]
y=X['k_cafe_tostado']
X = X.drop('k_cafe_tostado',axis=1)
```

Fig 22 datos de la clase modelo

```
pred_cafe_verde = model.predict(X_test)
#cafe tostado
X=datos[feature_cols]
y=X['k_cafe_tostado']
X = X.drop('k_cafe_tostado',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pTostion = model.score(X_test,y_test)
at = 1 - pTostion
ats = at *  100
pred_cafe_tostado = model.predict(X_test)
#color del cafe
X=datos[feature_cols]
y=X['color_pvolt']
X = X.drop('color_pvolt',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pcolor = model.score(X_test,y_test)
ac = 1 - pcolor
acs = ac * 100
pred_color = model.predict(X_test)

#temperatura del extractor
X=datos[feature_cols]
y=X['temp_ext1']
X = X.drop('temp_ext1',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pExtraccion = model.score(X_test,y_test)
ate = 1 - pExtraccion
ates = ate * 100

pred_temp_extraccion = model.predict(X_test)
#temperatura de la centrifuga
X=datos[feature_cols]
y=X['ext_temp']
X = X.drop('ext_temp',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pCentrifuga = model.score(X_test,y_test)
acn = 1 - pCentrifuga
acns = acn * 100


pred_temp_centrifuga = model.predict(X_test)
#efecto de evaporacion
X=datos[feature_cols]
y=X['ef2_producto']
X = X.drop('ef2_producto',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pEvaporacion = model.score(X_test,y_test)
aev = 1 - pEvaporacion
aevs = aev * 100

pred_eft_evaporacion = model.predict(X_test)
#brix de spray
X=datos[feature_cols]
y=X['ext_brix']
X = X.drop('ext_brix',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pSpray = model.score(X_test,y_test)
asp = 1 - pSpray
asps = asp * 100
```

Fig 23 datos de la clase modelo

```
X=datos[feature_cols]
y=X['ef2_producto']
X = X.drop('ef2_producto',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pEvaporacion = model.score(X_test,y_test)
aev = 1 - pEvaporacion
aevs = aev * 100

pred_eft_evaporacion = model.predict(X_test)
#brix de spray
X=datos[feature_cols]
y=X['ext_brix']
X = X.drop('ext_brix',axis=1)
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1, test_size=0.3)
model.fit(X_train,y_train)
pSpray = model.score(X_test,y_test)
asp = 1 - pSpray
asps = asp * 100

pred_brix_spray = model.predict(X_test)
```

In [42]:

```
cafe_verde = pred_cafe_verde
cafe_tostado = pred_cafe_tostado
color = pred_color
temp_extraccion = pred_temp_extraccion
temp_centrifuga = pred_temp_centrifuga
eft_evaporacion = pred_eft_evaporacion
brix_spray =pred_brix_spray
```

In [43]:

```
cvmin = cafe_verde.min()
ctmin = cafe_tostado.min()
colrmin = color.min()
tempextmin = temp_extraccion.min()
tempcenmin = temp_centrifuga.min()
eftevapmin = eft_evaporacion.min()
brixspmin = brix_spray.mean()
cvmax = cafe_verde.max()
ctmax = cafe_tostado.max()
colrmax = color.max()
tempextmax = temp_extraccion.max()
tempcenmax = temp_centrifuga.max()
eftevapmax = eft_evaporacion.max()
brixspmax = brix_spray.max()
```

In [44]:

```
print("\nPara la asignacion de los recursos de la calidad A/R se predice lo siguiente: ")
print("="*80)
print("\nPara TOSTION: \n")
#cafe arabica
if cvmin >= 185:
    print("="*65)
    print("- Se deben producir ","{:.2f}".format(ctmin) , " kilos de cafe Arabica por hora.")
    print("El cual tiene un equivalente a ","{:.2f}".format(cvmin), " kilos de cafe verde Arabica")
#tostion cafe robusta
    if cvmax <= 288:
        print("- Se deben producir ","{:.2f}".format(ctmax) , " kilos de cafe Robusta por hora.")
        print("El cual tiene un equivalente a ","{:.2f}".format(cvmax) , " kilos de cafe verde Robu
        print("y el color del cafe puede iniciar con un valor de ","{:.2f}".format(colrmin)  ,"y de

    else:
        print(cvmin,"Es un valor que no pertenece a la calidad A/R")
else:
    print(cvmax,"Es un valor que no pertenece a la calidad A/R")
```

Fig 24 datos de la clase modelo

```
    print("- Se deben producir ","{:.2f}".format(ctmin) , " kilos de cafe Arabica por hora.")
    print("El cual tiene un equivalente a ","{:.2f}".format(cvmin), " kilos de cafe verde Arabica")
#tostion cafe robusta
    if cvmax <= 288:
        print("- Se deben producir ","{:.2f}".format(ctmax) , " kilos de cafe Robusta por hora.")
        print("El cual tiene un equivalente a ","{:.2f}".format(cvmax) , " kilos de cafe verde Robu
        print("y el color del cafe puede iniciar con un valor de ","{:.2f}".format(colrmin)  ,"y de

    else:
        print(cvmin,"Es un valor que no pertenece a la calidad A/R")
else:
    print(cvmax,"Es un valor que no pertenece a la calidad A/R")
```

```
Para la asignacion de los recursos de la calidad A/R se predice lo siguiente:
================================================================================

Para TOSTION:

================================================================
- Se deben producir  67.32  kilos de cafe Arabica por hora.
El cual tiene un equivalente a  185.60  kilos de cafe verde Arabica
- Se deben producir  101.02  kilos de cafe Robusta por hora.
El cual tiene un equivalente a  287.40  kilos de cafe verde Robusta
y el color del cafe puede iniciar con un valor de  17.00 y debe mantenerce hasta  48.00
```

In [8]:
```
print("\nPara EXTRACCION")
print("="*35)
if tempextmin >= 120:
    print("="*65)
    print("- La temperatura inicial debe ser ","{:.2f}".format(tempextmin)  ,"grados")
    if tempextmax <= 142:
        print("  y debe llegar a","{:.2f}".format(tempextmax), "grados\n")
    else:
        print("{:.2f}".format(tempextmax),"la temperatura esta fuera del rango de la calidad A/R")
else:
    print("{:.2f}".format(tempextmin),"la temperatura esta fuera del rango de la calidad A/R")
```

```
Para EXTRACCION
===================================
====================================================================
- La temperatura inicial debe ser  121.10 grados
  y debe llegar a 142.00 grados
```

In [9]:
```
print("Para CENTRIFUGA")
print("="*35)
if tempcenmin >= 44:
    print("="*65)
    print("- La temperatura inicial debe ser ","{:.2f}".format(tempcenmin)   ,"grados")
    if tempcenmax <= 50:
        print("  y debe llegar a","{:.2f}".format(tempcenmax) , "grados\n")
    else:
        print("{:.2f}".format(tempcenmax) ,"la temperatura esta fuera del rango de la calidad A/R")
else:
    print("{:.2f}".format(tempcenmin) ,"la temperatura esta fuera del rango de la calidad A/R")
```

Fig 25 datos de la clase modelo

```
==================================
====================================================================
- La temperatura inicial debe ser  44.62 grados
  y debe llegar a 50.00 grados
```

```
]:
print("Para EVAPORACION")
print("="*35)
if eftevapmin >= 26:
    print("="*65)
    print("- El efecto inicial debe ser ","{:.2f}".format(eftevapmin)  )
    if eftevapmax <= 87:
        print("  y debe llegar a","{:.2f}".format(eftevapmax) ,"\n")
    else:
        print("{:.2f}".format(eftevapmin) ,"el efecto no pertenece a la calidad A/R")
else:
    print("{:.2f}".format(tempextmin) ,"el efecto no pertenece a la calidad A/R")
```

```
Para EVAPORACION
==================================
====================================================================
- El efecto inicial debe ser  26.59
  y debe llegar a 86.47
```

```
]:
print("Para SPRAY")
print("="*35)
if brixspmin >= 46:
    print("="*65)
    print("- El brix inicial debe ser ","{:.2f}".format(brixspmin))
    print("  y debe llegar a","{:.2f}".format(brixspmax) , "grados\n")

else:
    print("{:.2f}".format(brixspmin) ,"el brix no pertenece a la calidad A/R")
```

```
Para SPRAY
==================================
====================================================================
- El brix inicial debe ser  46.84
  y debe llegar a 55.06 grados
```

```
]:
print("Porcentajes en los que puede variar el valor de los recursos")
print("El cafe verde puede variar un","{:.2f}".format(avs),"%")
print("El cafe tostado puede variar un","{:.5f}".format(ats),"%")
print("El color del cafe puede variar un","{:.2f}".format(acs),"%")
print("La temperatura de extraccion puede variar un","{:.2f}".format(ates),"%")
print("La temperatura de la centrifuga puede variar un","{:.2f}".format(acns),"%")
print("El efecto de evaporacion puede variar un","{:.2f}".format(aevs),"%")
print("El brix de spray puede variar un","{:.2f}".format(asps),"%")
```

```
Porcentajes en los que puede variar el valor de los recursos
El cafe verde puede variar un 12.19 %
El cafe tostado puede variar un 0.00111 %
El color del cafe puede variar un 77.17 %
La temperatura de extraccion puede variar un 92.89 %
La temperatura de la centrifuga puede variar un 93.27 %
El efecto de evaporacion puede variar un 30.81 %
El brix de spray puede variar un 30.43 %
```

Fig 26 datos de la clase modelo

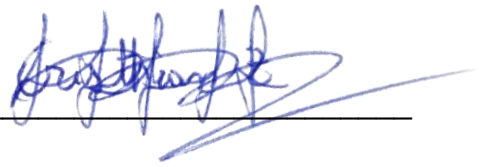
DECLARACIÓN Y AUTORIZACIÓN

Yo, **Zambrano Yont, Cristhian Oswaldo** con C.C: # 1311899692 autor del trabajo de titulación: **Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.** previo a la obtención del título de **Ingeniero en Sistemas Computacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, **13** de **mayo** de **2021**

f. 

Nombre: **Zambrano Yont, Cristhian Oswaldo**

C.C: 1311899692

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN			
TEMA Y SUBTEMA:	Diseño de un modelo predictivo, mediante la técnica de minería de datos para la asignación de recursos en la producción de café solido soluble para calidad A/R de la compañía ASKELGADO S.A.		
AUTOR(ES)	**Cristhian Oswaldo, Zambrano Yont**		
REVISOR(ES)/TUTOR (ES	**Ing. Castro Aguilar, Gilberto Fernando**		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	**Facultad de Ingeniería**		
CARRERA:	**Sistemas Computacionales**		
TITULO OBTENIDO:	**Ingeniero en Sistemas Computacionales**		
FECHA DE PUBLICACIÓN:	13 de mayo de **2021**	**No. DE PÁGINAS:**	100
ÁREAS TEMÁTICAS:	**Software, Minería de datos**		
PALABRAS CLAVES/ KEYWORDS:	Minería de datos, Python, Árbol de decisión, Modelo Predictivo		

RESUMEN/ABSTRACT (150-250 palabras): The present qualification work of the development of a design of a predictive model for the allocation of resources in the production of soluble solid coffee for A / R quality of the company ASKELGADO S.A. aims to identify the resources to be used, recognize a data mining technique, use a predictive model, and evaluate the model for A / R quality. The research used was of the documentary type, since an interview was conducted with the 2 people in charge of the production to find out their current status regarding the allocation of resources. For the project, the design of a predictive model was proposed, using the data mining technique to identify resources through historical data. To carry out the research, a qualitative, analytical research approach was used to study the context where the problem of resource identification exists; The decision trees technique are supervised learning techniques, in which functions are learned, relationships that associate inputs with outputs, so they conform to a set of examples of which we know the relationship between the input and the desired output. The mining technique is supervised, for that python will be used and the methodology is CRISP-MD. A tree was designed that generated a predictive evaluation model for decision making in the plant. In the end, the recommendations were made to be considered as the improvement of the previously used model

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR:	**Teléfono:** +593939097021	E-mail: cristhian.zambrano@cu.ucsg.edu.ec Cristhian2900@gmail.com
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UTE):	**Nombre: Ing. Edison, Toala Quimí, Mgs.**	
	Teléfono: +593-990976776	
	E-mail: edison.toala@cu.ucsg.edu.ec	
SECCIÓN PARA USO DE BIBLIOTECA		
Nº. DE REGISTRO (en base a datos):		
Nº. DE CLASIFICACIÓN:		
DIRECCIÓN URL (tesis en la web):		