UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

PROGRAMA DE ESTUDIO DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO EN COMPUTACIÓN Y SISTEMAS

Minería de datos y aprendizaje automático como soporte en la toma de decisiones en el área de producción y transporte de arándanos en la Empresa Agroberries s.a.c. – la libertad 2022

Línea de investigación:
Gestión de datos y de información

Autores:
Mendoza Vásquez, Jhordyn Daniel
Sánchez Otiniano, Luis Fernando

Jurado Evaluador:

Presidente: Ullón Ramírez, Agustín Eduardo
Secretario: Abanto Cabrera,Heber Gerson
Vocal: Castillo Robles, Edward Fernando

Asesor:
Lazo Aguirre, Walter Aurelio
Código ORCID: https://orcid.org/0000-0002-3223-4472

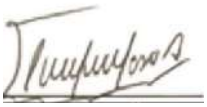
Trujillo–Perú
2023

Fecha de sustentación: 2023/07/21

MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022

por Mendoza Vasquez- Sanchez Otiniano

Fecha de entrega: 18-ago-2023 01:49a.m. (UTC-0500)
Identificador de la entrega: 2147442746
Nombre del archivo: 20230811-TESIS-FIX-Rev_13_ago.pdf (10.66M)
Total de palabras: 30946
Total de caracteres: 172752



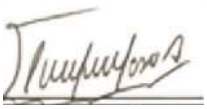
MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022

INFORME DE ORIGINALIDAD

3%	3%	0%	0%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.usil.edu.pe Fuente de Internet	2%
2	qdoc.tips Fuente de Internet	2%

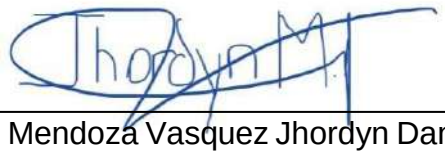


Excluir citas	Apagado	Excluir coincidencias	< 2%
Excluir bibliografía	Apagado		

ACREDITACIONES

MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022

Elaborado por:



Br. Mendoza Vasquez Jhordyn Daniel

Br. Sánchez Otiniano, Luis Fernando

Aprobado por

Ms. ULLON RAMIREZ, AGUSTIN EDUARDO
Presidente
CIP 137602

Ms. ABANTO CABRERA, HEBER GERSON
Secretario
CIP 106421

Ms. CASTILLO ROBLES, EDWARD FERNANDO
Vocal
CIP 192352

Dr. Lazo Aguirre, Walter Aurelio
Asesor

CIP 36034

UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

PROGRAMA DE ESTUDIO DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO EN COMPUTACIÓN Y SISTEMAS

Minería de datos y aprendizaje automático como soporte en la toma de decisiones en el área de producción y transporte de arándanos en la Empresa Agroberries s.a.c. – la libertad 2022

Línea de investigación:
Gestión de datos y de información

Autores:
Mendoza Vásquez, Jhordyn Daniel
Sánchez Otiniano, Luis Fernando

Jurado Evaluador:

Presidente: Ullón Ramírez, Agustín Eduardo
Secretario: Abanto Cabrera,Heber Gerson
Vocal: Castillo Robles, Edward Fernando

Asesor:
Lazo Aguirre, Walter Aurelio
Código ORCID: https://orcid.org/0000-0002-3223-4472

Trujillo–Perú
2023

Fecha de sustentación: 2023/07/21

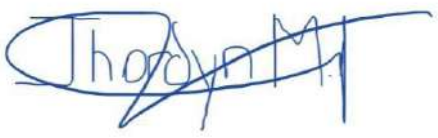
DECLARACION DE AUTENTICIDAD

Yo, Walter Aurelio Lazo Aguirre, docente del programa de Estudio de Pregrado de la Universidad Privada Antenor Orrego, asesor de la tesis “MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022”, de los autores Jhordyn Daniel Mendoza Vásquez y Luis Fernando Sanchez Otiniano, dejo constancia lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud del 3%. Así lo consigna el reporte de similitud emitido por el software Turnitin el día 18 de agosto de 2023.
- He revisado con detalle dicho reporte de la Tesis “MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022” y no se advierte índices de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las normas establecidas por la Universidad.

Trujillo, 18 de agosto de 2023

Lazo Aguirre Walter Aurelio
DNI: 17842663
ORCID: https://orcid.org/0000-0002-3223-4472



Mendoza Vasquez Jhordyn Daniel
DNI: 72857093

Sanchez Otiniano Luis Fernando
DNI: 75084813

DEDICATORIA

A MIS PADRES, HERMANA Y ABUELOS:

En gratitud a su dedicación, guía y confianza durante toda mi vida y en especial en aquellos momentos donde más los necesite.

Luis

A MIS PADRES Y HERMANOS:

Porque siempre mostraron en todo momento un gran interés y apoyo incondicional para cumplir con éxito mis objetivos, brindándome sus consejos y enseñanzas a lo largo de mi trayectoria profesional.

Jhordyn

AGRADECIMIENTO

A la Universidad Privada Antenor Orrego y a los docentes por la formación profesional que se nos brindó durante nuestra permanencia en la universidad.

Un agradecimiento especial al Dr. Walter Aurelio Lazo Aguirre por su asesoría e invalorable apoyo para la culminación exitosa del presente trabajo y a la Ing. Diana Flores por sus consejos y conocimientos que nos sirvió como base para nuestra tesis.

RESUMEN

MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022

Br. Mendoza Vásquez, Jhordyn Daniel
Br. Sánchez Otiniano, Luis Fernando

En el desarrollo del presente trabajo que se titula “Minería de datos y aprendizaje automático como soporte en la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022”, trata de la implementación de técnicas de minería de datos y de aprendizaje automático como árboles de clasificación, reglas de asociación y redes bayesianas, mismas que brindan un soporte para tomar decisiones a nivel estratégico al usuario final quienes serán los tomadores de decisiones.

El proceso de toma de decisiones, a través de la información que proporcionan las técnicas de minería de datos y de aprendizaje automático descritas anteriormente proveen mejoras a nivel de eficacia, eficiencia y satisfacción por parte de los tomadores decisiones.

Una vez se aplicaron las técnicas y se interpretó estadísticamente que resultaban confiables y usables para los intereses de la empresa, se lograron realizar reportes dinámicos, gráficos, rápidos y que a su vez son eficaces y del agrado de los tomadores de decisiones.

Palabras clave: Minería de datos, Técnicas de Data Mining, Toma de decisiones, Aprendizaje automático.

ABSTRACT

DATA MINING AND MACHINE LEARNING AS SUPPORT IN DECISION-MAKING IN THE AREA OF BLUEBERRY PRODUCTION AND TRANSPORTATION AT THE COMPANY AGROBERRIES S.A.C. – LA LIBERTAD 2022

Br. Mendoza Vásquez, Jhordyn Daniel
Br. Sánchez Otiniano, Luis Fernando

In the development of this work entitled "Data mining and machine learning as support for decision making in the area of production and transportation of blueberries in the company Agroberries S.A.C. La Libertad 2022", it is about the implementation of data mining techniques and machine learning as classification trees, association rules and Bayesian networks, which provide support for decision making at the strategic level to the end user. - La Libertad 2022", deals with the implementation of data mining and machine learning techniques such as classification trees, association rules and Bayesian networks, which provide support for strategic decision making to the end user who will be the decision makers.

The decision making process, through the information provided by the data mining and machine learning techniques described above, provides improvements in terms of effectiveness, efficiency and satisfaction on the part of the decision makers.

Once the techniques were applied and it was statistically interpreted that they were reliable and usable for the company's interests, it was possible to create dynamic, graphic and fast reports that are effective and to the liking of the decision makers.

Keywords: Data Mining, Data Mining Techniques, Decision Making, Machine Learning.

PRESENTACIÓN

Señores Miembros del Jurado:

Cumpliendo con los requerimientos estipulados en el reglamento de Grados y Títulos de la Facultad de Ingeniería de la Universidad Privada Antenor Orrego, para obtener el Título de Ingeniero de Computación y Sistemas, pongo a vuestra disposición la presente tesis titulada: **APLICACIÓN DE MINERÍA DE DATOS COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022.**

Gracias.

Trujillo, 13 de julio de 2023

Br. Mendoza Vásquez, Jhordyn Daniel

Br. Sánchez Otiniano, Luis Fernando

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS Y GRÁFICOS

CAPÍTULO I:
INTRODUCCIÓN

CAPÍTULO I: INTRODUCCIÓN

1.1. PROBLEMA DE INVESTIGACIÓN

1.1.1. REALIDAD PROBLEMÁTICA

Una de las razones por las que las empresas logran posicionarse competitivamente en posiciones privilegiadas en sus determinados campos de negocio, es la medida en la que sus procesos internos – externos están alineados con la tecnología y cómo aprovechan los tomadores de decisiones empresariales este alineamiento haciendo uso de herramientas tecnológicas, metodologías, entre otros artefactos, para el beneficio de la organización.

Uno de estos procesos a los cuales se le saca provecho es la minería de datos, sin embargo, tiene limitaciones dentro de los contextos empresariales PYME, ya que, por la carencia cultural a nivel tecnológico y el desconocimiento total sobre cómo aplicar una minería de datos; es importante en este punto hacer mención sobre las metodologías que existen y, que, dentro de un proceso especificado, mantienen a la denominada Data Mining como un proceso clave para el logro de objetivos. Estas metodologías pueden ser PROCESO KDD, CRISP-DM, SEMMA, etc.

En Perú, se logra escuchar poco sobre la minería de datos aplicada bajo un contexto metodológico y tampoco se logra notar investigaciones acerca de este tema. Si excluimos a trabajos basados en empresas comerciales retail y que son ampliamente conocidas como: PLAZA VEA, METRO, TOTTUS, entre otras, nos quedamos con un escaso margen de exploración sobre el Data Mining y las metodologías quedan casi en una nula intención de descubrimiento.

En La Libertad, en palabras de (Agromedo Cueva & Salazar Ávila, 2019): “La mayoría de empresas industriales del departamento de La Libertad no usan herramientas tecnológicas ni inteligencia de negocios para dar apoyo o si quiera soporte a la toma de decisiones, pero es de vital importancia contar con estas para que así las empresas tanto públicas como privadas tengan un principal recurso para la toma de decisiones en la alta dirección”.

En la empresa Agroberries S.A.C, que está identificada con RUC 20600807685, se ubica en Virú que pertenece a la región La Libertad; presenta problemas en el área de producción y transporte además del desconocimiento que existe con relación a los datos y la información que los mismos puedan contener; en específico, existen sobrecostos en las cosechas y es que se plantean cada año presupuestos de producción por

kilo de arándano cosechado, pero en muchas ocasiones este se ve sobre alzado y se desconoce el motivo. Luego, los costes de transporte hacia los recintos de compradores (también llamados Kias) tienen un desbalance en el porcentaje de cumplimiento con respecto al transporte de kilos de arándanos; por último y como bien se ha especificado líneas anteriores, no se tiene información ni conocimiento alguno sobre sobre los datos y las relaciones en los reportes diarios, mismos que incluyen datos como los kilos cosechados, número de cosechadores, fechas, fundos, etc. Con todo lo anterior, los tomadores de decisiones se ven envueltos en una problemática que abarca a la no disponibilidad de data sólida para una correcta toma de decisiones que a su vez influye en los resultados finales como las cosechas, utilidades y desenvolvimiento de procesos.

1.1.2. ENUNCIADO DEL PROBLEMA

¿En qué medida la aplicación de la minería de datos y aprendizaje automático sirve como herramienta de soporte para la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022?

1.1.3. ALCANCE

La presente tesis tiene como alcance aplicar ciencia de datos en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C., misma que pertenece al sector agrícola comercial en el departamento de La Libertad; del modo tal que permita determinar si ha sido favorable para la empresa la aplicación del Data Mining y Machine Learning, y así captar los problemas resueltos a través de las técnicas propuestas.

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

Comprobar que la aplicación de la minería de datos y aprendizaje automático sirve como herramienta de soporte para la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022.

1.2.2. OBJETIVOS ESPECÍFICOS

- OE1: Analizar la data del área de producción y transporte de arándanos de la empresa Agroberries, para formar un dataset compuesto solo por aquellas características a usar.
- OE2: Implementar técnica de árbol de decisión para predecir con base en el sobrecosto.
- OE3: Implementar Binning y reglas de asociación con fin de hallar las relaciones entre valores agrupados con respecto al cumplimiento de Kias.
- OE4: Implementar redes bayesianas para predecir el cumplimiento de la cantidad de arándanos cosechados con base en características del detalle diario haciendo uso del software SPSS Modeler.
- OE5: Visualización de los resultados estadísticos a través de gráficos obtenidos por cada software usado.

1.3. JUSTIFICACIÓN DEL ESTUDIO

La data que generalmente es percibida como “no importante” es aquella que esconde patrones que pueden ser muy bien aprovechados por las empresas en sus distintos rubros; desde empresa dedicada a la agricultura hasta un mini-market. Las conexiones o patrones interesantes no soy vitos a simple vista, sino que tienen que ser procesados haciendo minería de datos para descubrir lo relevante y aprovechar aquella información para beneficio de la organización, no importa el tamaño de esta; sin embargo, es de conocimiento popular que las empresas grandes y trasnacionales, son las que más aprovechan esta data debido a muchos factores, entre ellos el costo que implica, la inversión de tiempo, entre muchos otros.

En las regiones del Perú no existe una gran cantidad de información con respecto a la aplicación de minería de datos en empresas del sector agrícola y esto debido a que las mismas no se interesan o no tiene un mínimo cultura tecnológica como para tomar la iniciativa de emprender e invertir en el descubrimiento de conocimiento que no se sabe y captar las oportunidades de mejora gracias a información “escondida”, pero relevante. En La Libertad, por ejemplo, es una región sin exploración de minería de datos a grandes rasgos; de hecho, que existe una variedad de trabajos, pero enfocados al ámbito de empresa grande y/o transnacional, por colocar un ejemplo, del sector retail, como bien puede ser Plaza Vea, Tottus, Metro, entre muchas otras y se complica cuando se quiere explorar otros sectores debido a que las empresas son celosas en demasía cuando se trata de compartir su data.

El proyecto tiene como justificación brindar el conocimiento a la región de qué en cualquier empresa que cuente con data registrada puede tener una oportunidad de crecimiento con base en la minería de datos y aprendizaje automático y el descubrimiento de patrones, que de por sí también incluye una iniciativa a formar cultura tecnológica entre los integrantes de una compañía.

CAPÍTULO II:
MARCO DE REFERENCIA

CAPÍTULO II: MARCO DE REFERENCIA

2.1. ANTECEDENTES DEL ESTUDIO

2.1.1. (GRÁNDEZ, 2018), en su trabajo, **"Aplicación de la minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales"**, tienen como objetivo analizar los datos y determinar tendencias en la compra de suplementos nutricionales tomando en cuenta variables como la edad, el género, el peso, la estatura, estado civil, número de hijos, ocupación, ingreso mensual e incluso la actividad física; luego de aplicar la técnica de Clustering se obtuvo las tendencias más notorias dando cierto margen de confianza para predecir las tendencias de compra con el modelo usado.

Los investigadores indican que, dentro de la empresa, hasta antes de tomarla como modelo, no se analizaron datos para medir tendencias de compra ni para predecir qué características de los clientes determinan la compra de un producto.

De forma exacta y precisa consideran a las siguientes variables:

- Edad, genero, peso, estatura, estado civil.
- Actividad física.
- Ocupación, ingreso mensual.
- Número de hijos.

La muestra que tomaron fue de 611 clientes que realizaron compras entre enero del 2016 a junio del mismo año.

La técnica de minería de datos aplicada en el trabajo de investigación fue Clustering haciendo uso del algoritmo proporcionado por Microsoft que a su vez incluye técnicas de asociación de la empresa previamente mencionada.

Las herramientas que se usan en la investigación fueron:

- Microsoft SQL Server, con el cual pudieron generar base de datos transaccionales.
- Microsoft Integration Services en la versión del año 202, de esta manera pudieron hacer uso de la extensión Analysis Services para el proceso de minería de datos.

El objetivo general que trata la tesis es la de aplicar la minería de datos haciendo uso de un software informático, en este caso Integration Services de Microsoft para determinar los patrones de consumo futuro en la distribuidora de suplementos nutricionales "Lab Nutrition" con la finalidad de implementar políticas que incrementen el nivel de ventas.

Después de ello hace hincapié en los objetivos específicos se desglosarán entre la identificación de reglas de consumo con base a las características del consumidor (estas variables han sido descritas anteriormente); luego, la formación de paquetes de productos de acuerdo a reglas de consumo, posterior a ello identificación de los productos que prefieren los clientes y aplicar el Cros-Selling, así finalizar con la demostración de la probabilidad y la importancia de las reglas que determinan el consumo.

Lo más interesante de la tesis para los intereses del presente trabajo de investigación y la razón por la que fue elegido como antecedente, es que el autor puede generar una base de datos transaccional haciendo uso de la herramienta Microsoft SQL Server a partir de un archivo plano que es compartido por la empresa dedicada a la venta de suplementos nutricionales; esta es una manera sencilla de dar inicio al trabajo de minería de datos, ya que, en otros contextos de investigación se puede dar el caso de que la data venga encasillada en tablas dinámicas sobre otras tablas dinámicas. Este antecedente no presente tales inconvenientes por lo que es más factible la manipulación inicial de sus datos.

Al finalizar el proceso de minería de datos haciendo uso de las herramientas mencionadas, pueden ver la luz los resultados esperados que se grafican en las siguientes tablas con sus determinadas reglas que muestran el patrón de los consumidores.

Tabla 1: Patrones de consumo para producto HARBINGER FITNESS

Patrones de consumo en los clientes del producto denominado HARBINGER FITNESS CORREA BIG GRIP PRO LIFTING STRAPS			
	Probabilidad	**Importancia**	**Regla**
1	1,000	2,45	▪ Actividad física = PILATES, ▪ Cliente edad = 25-32
2	1,000	2,45	▪ Actividad física = PILATES, ▪ Cliente ingreso mensual = 4960,7 – 6771,2
3	1,000	2,45	▪ Actividad física = PILATES, ▪ Cliente estatura ≥ 1,78
4	1,000	2,45	▪ Actividad física = PILATES, ▪ Cliente peso = 73,52 – 80,48

Tabla 2: Patrones de consumo para producto SYNTRAX NECTAR

Patrones de consumo en los clientes del producto denominado SYNTRAX NECTAR MEDICAL			
	Probabilidad	**Importancia**	**Regla**
1	1,000	0,52	▪ Cliente número hijos >=3, ▪ Actividad física = BICICLETA ESTÁTICA (SPINNING)
2	1,000	0,48	▪ Cliente número hijos >=3, ▪ Cliente estatura < 1,72
3	1,000	0,45	▪ Cliente número hijos >=3, ▪ Cliente peso = 80,4 – 91,5
4	1,000	0,39	▪ Cliente número hijos >=3, ▪ Actividad física = CAMINAR
5	1,000	0,39	▪ Cliente número hijos >=3, ▪ Cliente estado civil = S
6	1,000	0,39	▪ Cliente peso >=91,49788672, ▪ Cliente sexo = F
7	1,000	0,39	▪ Cliente edad < 25, ▪ Cliente ingreso mensual = 6771,2 – 8184,7

Tabla 3: Patrones de consumo para producto MUSCLETECH PLANITUM 100%

Patrones de consumo en los clientes del producto denominado MUSCLETECH PLATINUM 100% WHEY			
	Probabilidad	**Importancia**	**Regla**
1	1,000	1,31	▪ Actividad física= ELÍPTICA, ▪ Cliente ingreso mensual >= 9731
2	1,000	1,31	▪ Cliente peso >=91,5 ▪ Actividad física = CAMINAR
3	1,000	1,31	▪ Actividad física = PESAS, ▪ Cliente ingreso mensual < 4960,7

4	1,000	1,31	▪ Actividad física = PESAS, ▪ Cliente edad = 25 – 32
5	1,000	1,31	▪ Actividad física = PESAS, ▪ Cliente estatura = 1,74 – 1,78
6	1,000	1,31	▪ Actividad física = PESAS, ▪ Cliente peso = 65,39052120 – 73,5
7	1,000	1,31	▪ Actividad física = PESAS, ▪ Cliente estado civil = S

También usan modelos neuronales, los cuales presentan los siguientes resultados por cada producto del mismo modo que las reglas.

Tabla 4: Resultados modelo neuronal MUSCLETECH HYDROXYCUT PRO

Resultados modelo neuronal para el producto MUSCLETECH HYDROXYCUT PRO CLINICAL GUMMIES			
Item	**Atributo**	**Valor**	**Probabilidad de consumo**
1	CLIENTE NÚMERO DE HIJOS	>= 3	100
2	CLIENTE ESTATURA	< 1,72	65,71
3	ACTIVIDAD FÍSICA	BAILAR	64,49
4	CLIENT EDAD	>= 43	55,22
5	CLIENTE NÚMERO DE HIJOS	< 2	39,83
6	CLIENTE ESTADO CIVIL	C	35,08
7	ACTIVIDAD FÍSICA	NADAR	28,78
8	ACTIVIDAD FÍSICA	MONTAR EN BICICLETA	28,04
9	CLEINTE PESO	< 65	23,69
10	ACTIVIDAD FÍSICA	CAMINAR	22,94
11	CLIENTE EDAD	25 – 32	20,07

12	CLIENTE EDAD	32 – 39	19,96
13	ACTVIDAD FÍSICA	PILATES	10,47

Tabla 5: Resultados modelo neuronal NUTREX LIPO-6 BLACK

Resultados modelo neuronal para el producto NUTREX LIPO-6 BLACK			
Item	**Atributo**	**Valor**	**Probabilidad de consumo**
1	ACTIVIDAD FÍSICA	PESAS	81,97
2	CLIENTE NRO DE HIJOS	>= 3	62,46
3	CLIENTE ESTATURA	< 1,72	56,59
4	CLIENTE ESTATURA	1,74 – 1,78	53,91
5	ACTIVIDAD FÍSICA	ELÍPTICA	51,13
6	ACTIVIDAD FÍSICA	KICKBOXING	48,77
7	CLIENTE INGRESO MENSUAL	< 4961	23,91
8	CLIENTE PESO	< 65	21,91
9	CLIENTE ESTADO CIVIL	C	17,77
10	CLIENTE EDAD	39 – 43	13,96
11	ACTIVIDAD FÍSICA	PILATES	10,12
12	CLIENTE EDAD	< 25	5,56
13	CLIENTE EDAD	>= 43	3,28
14	CLIENTE PESO	>= 91	1,23

Al final de la tesis se redactan las conclusiones que son las siguientes:

- Es muy importancia la metodología que se haya seleccionado para realizar el trabajo de investigación, de esta manera y, tras

una breve comparación entre metodologías, se optó por CRISP-DM.
- Del mismo nivel de importancia, la selección de algoritmos adecuados es de suma relevancia para conseguir los resultados trazados.
- Encontraron el producto más demandado y se podría predecir el futuro de cuál será el producto más demandado tras distintas variables que tienen la posibilidad de ser introducidas en la minería de datos.
- Los resultados obtenidos pueden aplicarse a otros locales que cumplan similares características a nivel social y económico como del contexto en el que estén su origen de datos.

2.1.2. (Flores, 2021), en su trabajo, **"Clasificación de cultivos de quinua orgánica mediante el uso de imágenes aéreas multiespectrales y técnicas de aprendizaje automático"**, tienen como objetivo usar árboles de decisión, análisis discriminante, máquinas de vectores de soporte, K-means, clasificadores de conjuntos y métodos de aprendizaje profundos para mapear los cultivos de quinua haciendo uso de todo lo mencionado anteriormente y clasificarlos a partir de imágenes con la finalidad de resolver la gestión agrícola y la seguridad alimentaria; para ello se las imágenes que se toman son de un sistema aéreo no tripulado y son multiespectrales. Al final del trabajo se logra obtener la fidelidad suficiente para usar los métodos de clasificación en el mapeo de cultivos de quinua y son útiles para predecir los rendimientos.

El antecedente que se muestra a continuación toma como campo de estudio a los campos de cultivo de quinua en Cabana, que se ubica en la región de Puno; como bin se ha descrito anteriormente, las técnicas que se usan son varias y se dividen en:

- SVM.
- K-means.
- Análisis discriminante.
- Clasificadores conjuntos.
- Deep-Learning.
- Árboles de decisiones.

Son justamente los últimos a los cuáles le asignamos mayor importancia debido a que es la razón por la que tomamos el trabajo de investigación descrito para compararlo con nuestros objetivos específicos.

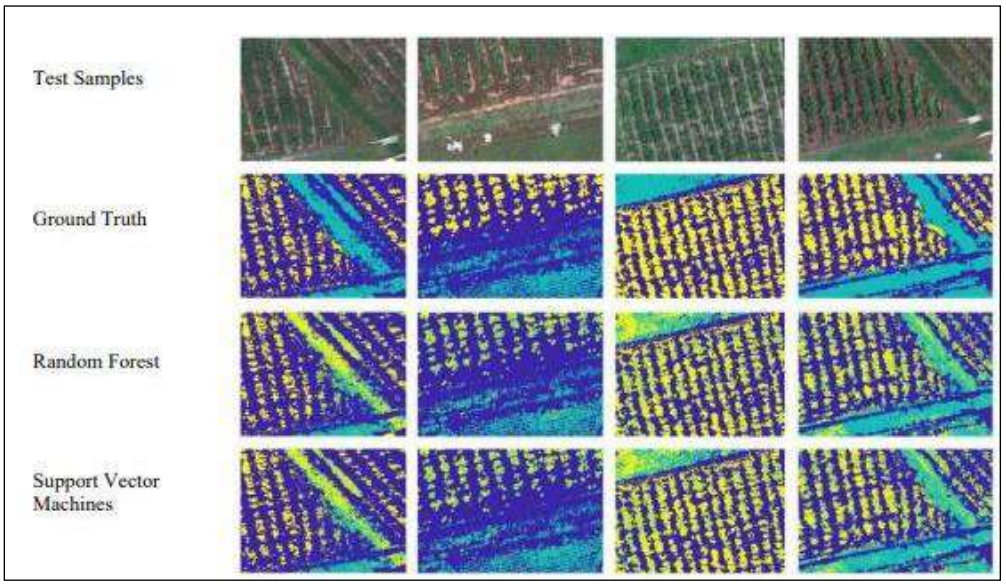
La tesis también presente una problemática y esta se centra en la gestión agrícola y la seguridad alimentaria que se ocasiona debido al escaso uso de tecnologías productivas, hay otros factores como el incremento de plagas, la variabilidad cada vez es más imprescindible de condiciones climáticas lo cual genera un incremento en niveles de

pobreza y pobre extra; lo que se requiere es que al mapear los cultivos sirvan como herramienta para subsanar estos problemas que vienen ocurriendo.

El autor indicia que para la solución a ello se evalúan diversos métodos de clasificación – los descritos anteriormente – para de esta manera clasificar el cultivo de quinua y así el mapeo será de manera automática, haciendo uso de las técnicas de Machine Learning colocadas anteriormente. Esto se realizará a partir de imágenes en áreas multiespectrales tomadas desde un dron o un sistema aéreo no tripulado.

A continuación, se puede apreciar las fotos que se tomaron desde el dron.

Imagen 1: Fotografías en el área multiespectral del cultivo de quinua



Fuente: Flores, 2021

Para iniciar el proceso de aplicar técnicas de clasificación se recogieron, luego de una segmentación semántica haciendo uso de redes neuronales convolucionales, se pudo capturar a través de las imágenes anteriormente mostradas, datos que servirán, en su conjunto, como un dataset para el inicio de la experimentación.

Tabla 6: Dataset del número de píxeles para entrenamiento

DATASET PARA ENTRENAMIENTO NÚMERO DE PÍXELES				
Data	**Date**	**No vegetación**	**Otra vegetación**	**Quinua**
1	27-12-2018	2929509	159406	664265
2	04-01-2019	3923729	1251091	593028
3	25-01-2019	1365821	3330347	203280
4	02-02-2019	1339181	1202338	236851
5	16-02-2019	1659807	3594423	268550
6	28-02-2019	1628694	2588546	135650
Total		12846741	12126151	2101624

Una vez explicado el proceso en el que consiste la tesis tomada como antecedente y luego de haber empleado las técnicas de clasificación haciendo uso del dataset mostrado, podemos mostrar los resultados en

cuanto a precisión de los modelos usado se refiere, obteniendo la siguiente tabla.

Tabla 7: Precisión de los modelos clasificadores

Clasificador	Precisión
Árbol de decisión	91,2%
Análisis discriminante	91,8%
SVM	95,4%
KNN	94,5%
Random Forest	94,7%
Adaboost	91,2%

Como una conclusión final general de la tesis podemos decir que gracias a los métodos de clasificación que se usaron y la precisión obtenida, recalcado que no se tomó en cuenta el aprendizaje profundo también llamado “Deep Learning”, se puede determinar un alto grado de fidelidad para mapear cultivos de quinua que pueden ser usados como etapa previa para la predicción en los rendimientos y una solución eficaz y efectiva a los problemas que se están presentando de la gestión agrícola y la seguridad alimentaria. De este modo, se puede afirmar que se ha cumplido el objetivo de la tesis con éxito.

2.1.3. (Araujo, 2018), en su trabajo, **“Reglas de asociación y predicción utilizando series de tiempo”**, tienen como objetivo el apoyo a las actividades productividad de pequeños productores agrícolas haciendo uso de minerías de datos; de esta manera el autor crea un módulo de análisis de datos usando reglas de asociación con el algoritmo “A priori” para así adicionando en el sistema actual que contiene varios otros circuitos de comercialización, es importante recalcar como último punto que se hace uso de series de tiempo (fechas) de tal modo que se establece un periodo de inicio y fin lo que también afecta a la confianza y otras variables estadísticas usadas en las reglas de asociación como el Lift.

Tal cual se ha descrito previamente, este trabajo tiene como objetivo apoyar a las actividades de producción de los pequeños agricultores, para ello describe a sus objetivos específicos como un desglosamiento de pequeñas actividades para lograr alcanzar con éxito el objetivo general, de esta manera se divide en:

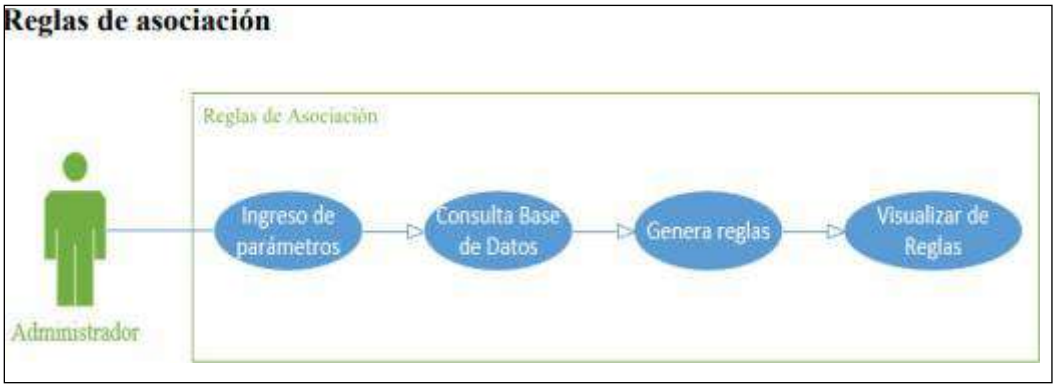
- El establecimiento de los productos de mayor consumo con base en los algoritmos de reglas de asociación y del mismo modo para las series de tiempo.
- Luego, establecer un procedimiento que permita al autor analizar datos desde una hoja de Excel para un posterior análisis.

- Tras ello, crear un módulo de análisis de datos con la opción de reglas de asociación haciendo uso del algoritmo “A priori”.
- Finalmente, crear en el mismo módulo de análisis de datos la opción para predecir usando series de tiempo.

La tesis pone marcha al diseño y usa el framework MVC en donde define modo modelo a la base de datos de las ventas que se hacen en el ministerio de agricultura y como vista a los componentes actuales en ese momento que es un esquema que funciona con tecnología PrimeFaces, como último punto el controlador se basa en los parámetros de las reglas de asociación que en este caso es la confianza, el soporte mínimo y el tiempo para lo que es series de tiempo, esto y los registros consultados se generará las reglas de asociación.

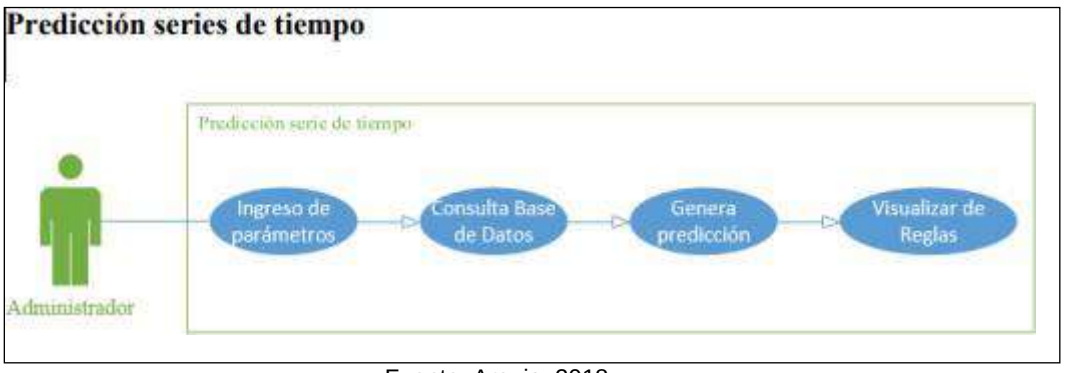
El autor usa la metodología SCRUM para el proyecto y por tanto realiza el respectivo product backlog y el sprint backlog acompañado de las historias de usuario. Tras ello también realiza diagramas de caso de uso para reglas de asociación y la predicción que se basará en las series de tiempo, así fueron sus resultados.

Imagen 2: Diagrama casos de uso para reglas de asociación



Fuente: Araujo, 2018

Imagen 3: Diagrama casos de uso para predicción de series de tiempo



Fuente: Araujo, 2018

Es importante saber que además de crear una interfaz para la puesta en marcha de las reglas de asociación y la predicción basándose en series de tiempo, también tienen como una meta lograr que la interfaz pueda leer parámetros de una hoja de Excel, de esta manera realizan el prototipo de la interfaz quedando de la siguiente forma.

Imagen 4: Prototipo interfaz del sistema de información

Fuente: Araujo, 2018

El punto clave de la tesis y por la cual la escogimos como antecedente es cuando, tras haber implementado la interfaz haciendo uso de todo lo dicho previamente, se obtiene como resultado las reglas de asociación y se compara con el software Weka, del tal modo se puede observar que los resultados del sistema SINARED son parecidos a los del software WEKA.

Tabla 8: Comparación entre resultados SINARED vs WEKA

COMPARATIVA RESULTADOS SINARED vs WEKA		
#	**SINARED**	**WEKA**
1	COL MORADA = t22 → MELLOCO = t22	COL MORADA = t22 → MELLOCO = t22
2	RABANO = t22 → COL MORADA = t22	RABANO = t22 → COL MORADA = t22
3	COL MORADA = t22 → RABANO = t22	COL MORADA = t22 → RABANO = t22
4	REMOLACHA = t22 → COL MORADA = t22	REMOLACHA = t22 → COL MORADA = t22

5	COL MORADA = t22 → REMOLACHA = t22	COL MORADA = t22 → REMOLACHA = t22
6	ESPINACA = t22 → MELLOCO = t22	ESPINACA = t22 → MELLOCO = t22
7	RABANO = t22 → MELLOCO = t22	RABANO = t22 → MELLOCO = t22
8	REMOLACHA = t22 → MELLOCO = t22	REMOLACHA = t22 → MELLOCO = t22
9	REMOLACHA = t22 → RABANO = t22	REMOLACHA = t22 → RABANO = t22
10	RABANO = t22 → REMOLACHA = t22	RABANO = t22 → REMOLACHA = t22

Por consiguiente, los puntos con los que concluye el autor son los siguientes:

- Los resultados arrojados por el sistema SINARED son los mismos que arroja el software WEKA, de esta manera se puede deducir que las reglas de asociación que han sido implementadas en la tesis se encuentran correctamente definidas.

- Corroborar un hecho; las reglas de asociación conceden la opción de conocer resultados tanto generales como específicos, lo que quiere decir, que se pueden visualizar resultados de las reglas más importantes entre distintos productos.

2.1.4. (Calvo-Valverde et al., 2019), en su trabajo, **"Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka negra y la productividad en cultivos agrícolas"**, tienen como objetivo demostrar si las redes bayesianas tiene la capacidad para predecir de forma adecuada o con un alto índice de confianza o precisión el avance de la plaga Sigatoka negra y la productividad del cultivo de banano haciendo uso de datos proporcionados por la empresa CORABANA referidos al clima histórico. Se obtiene como conclusión que las redes bayesianas dinámicas o RBDs no tiene la capacidad suficiente para predecir el avance de la plaga y por tanto no es una técnica fiable; esto se obtiene de su grado de precisión que se consigue del modelo final.

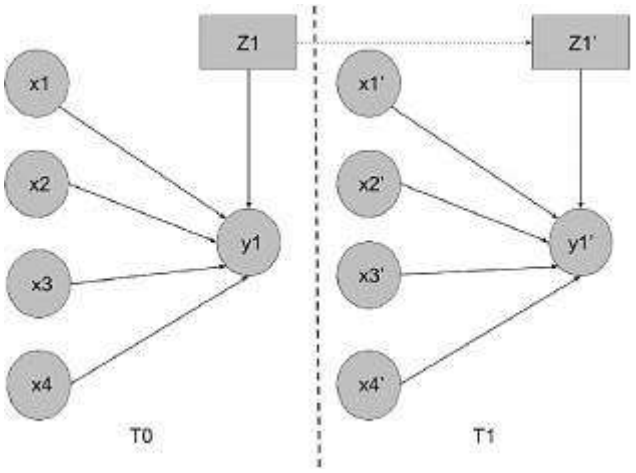
Este es un paper que nos resultó interesante por la incapacidad que resulta de evaluar a las redes bayesianas dinámicas como soporte para la predicción en el avance de la Sigatoka negra y la productividad en los cultivos agrícolas.

Lo primero que realizan los autores es una revisión a los antecedentes y trabajadores relacionados, para después introducirnos al diseño experimental de la investigación que inicia con la obtención de los

datos por la empresa CORABANA referente a la data histórica del clima y el cultivo de banano; después diseñan e implementan las redes bayesianas dinámicas y una red bayesiana que representará las relaciones encontradas en los datos,

A continuación, se muestra el diseño de la red bayesiana dinámica propuesta por los autores.

Imagen 5: Diseño de RBD

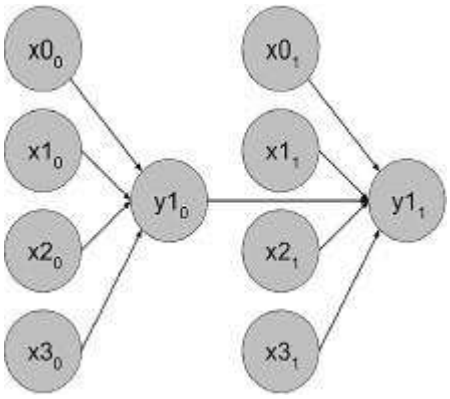


Fuente: Calvo-Valverde et al., 2019

Los autores definen a su diseño de red bayesiana dinámica de la siguiente manera: “Los arcos entre los nodos de la red representan que un nodo puede influenciar a otro nodo de la red. La influencia fluye a través de la red en los casos en los que los caminos entre los nodos a través de los arcos que los unen sean caminos activos”.

El siguiente diseño es el de una red bayesiana como tal, esta misma representa a una red bayesiana dinámica para dos ventanas, es decir, desarrollada en dos tiempos específicamente en el intervalo de dos semanas.

Imagen 6: Diseño de RB



Fuente: Calvo-Valverde et al., 2019

Los autores definen a su red bayesiana de la siguiente manera: “Cada nodo tiene asociada una tabla de distribución condicional que representa la probabilidad de su valor dado el valor de cada uno de sus padres. La información en este tipo de red fluye ya que cada nodo de salida es padre del nodo de salida en la siguiente ventana, influyendo así su valor”.

Las conclusiones de este trabajo es lo más interesante y es la razón por la cual fue escogida como antecedente para la presente investigación ya que en la efectividad de las redes bayesianas se obtuvo una efectividad de predicción realmente baja y es que para el conjunto micro llegó a un 0,49 en el contexto más favorable, mientras que el contexto macro llegó solamente a 0,16, por lo que no es posible contar como soporte a estas redes bayesianas dinámicas. Por lo que el autor recomienda hacer uso de las Robs (redes bayesianas) incluso si se tienen que trabajar con varias ventanas, mismas que habíamos visto representan al intervalo de tiempo con el cual se experimentará.

2.1.5. (Zamora, 2018), en su trabajo**, "Aplicación de técnicas de minería de datos para pronósticos del sector agrícola"**, tiene como objetivo encontrar un modelo que sea capaz de pronosticar la producción y superficie sembrada de cultivos agrícolas tales como la papa y el trigo de algunas zonas determinadas que se explican en detalle en el informe. Usa varias técnicas como series de tiempo, regresión, SVM y redes neuronales; a pesar de que se concluye que para el trabajo no es eficaz ninguna de los modelos empleados, se plantea el caso de que se tuviera que escoger uno de los cuatro, se optaría por el de las redes neuronales y se plantea su presentación de resultados correctamente interpretables a través de gráficos.

El autor comienza explicando y contextualizando al trabajo de investigación en el negocio en el que se encuentran e indica que se trata, de forma general, de la agricultura de Chile y sigue ahondando en que la papa y el trigo forman parte de los cultivos de mayor importancia en el mundo acompañados del maíz y del arroz.

De esta manera plantea el problema que tratará la investigación, que se centra en la evolución constante de la reducción de siembra y que esta no está a la par con todos los cultivos sino del trigo, remolacha, papas, raps y leguminosas; en su sentido opuesto tenemos a los rendimientos por hectárea y que han aumentado en todos los cultivos e incluso han llegar a topes de niveles importantes. Es este el problema por el cual se requiere aplicar técnicas de minería de datos para pronosticar o estimar la superficie sembrada y cosecha de cultivos anuales.

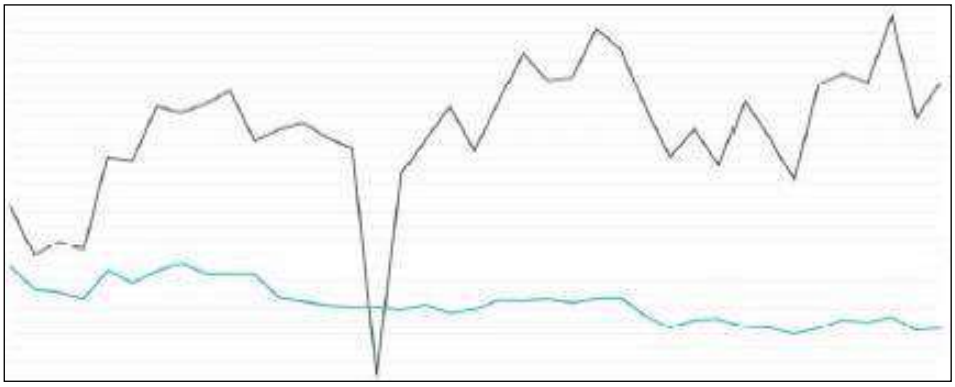
Primero, el autor inicia con el desarrollo del estado de arte, tras ello determina las técnicas de minería de datos que se van a emplear y toma en consideración aquellas que permitirán obtener la mejor proyección en la producción de determinados productos; luego evalúan los modelos usado y comparan los resultados para indicar qué tipo de modelo es los suficientemente capaz para pronosticar lo que se requiere, es decir, generar confianza para basarse en las predicciones con respecto a la producción en ciertos productos.

Antes del inicio del desarrollo a la propuesta de solución se destaca la selección de metodología, la cual será CRISP-DM y tras ello la herramienta a usar que es WEKA en su versión 3.8; seleccionada en muchos casos por ser de código abierta y gratuita, este trabajo de

investigación no es diferente, ya que, son los puntos para considerar para su elección.

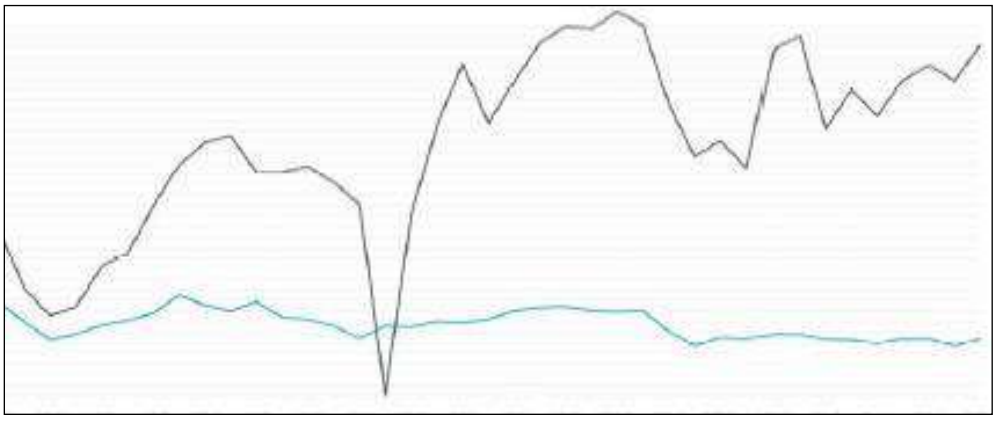
Lo que más se destaca de este antecedente y por el cuál ha sido seleccionado es por sus visualizaciones gráficas tanto en la parte de análisis de datos como después de haber implementado las técnicas de minería de datos. A continuación, podemos observarlas de forma que podamos darnos una idea de la interpretación.

Imagen 7: Superficie y producción de Trigo en El Biobío



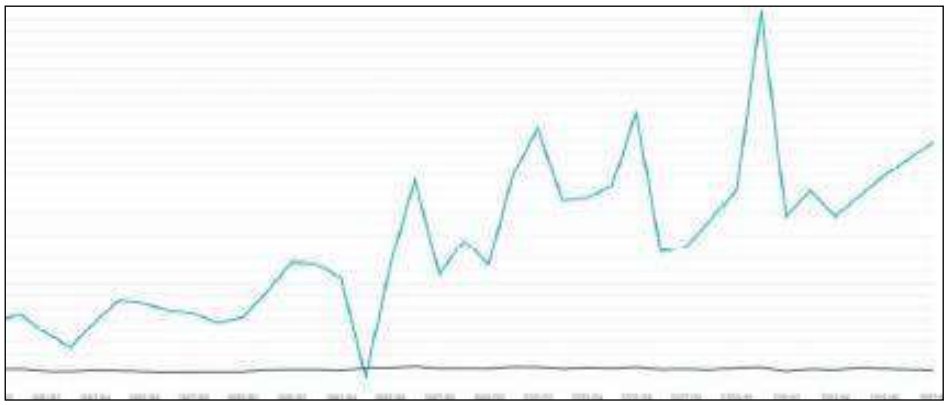
Fuente: Zamora, 2018

Imagen 8: Superficie y producción de trigo en La Araucanía.



Fuente: Zamora, 2018

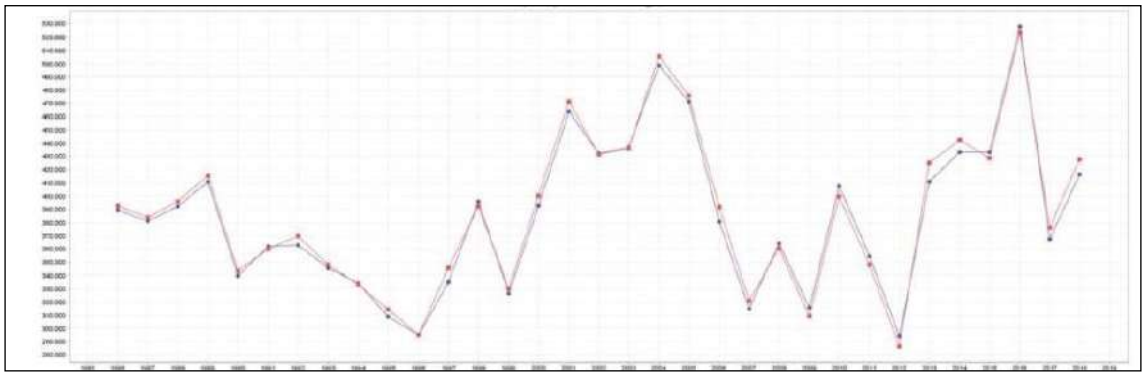
Imagen 9: Superficie y producción de papa en La Araucanía



Fuente: Zamora, 2018

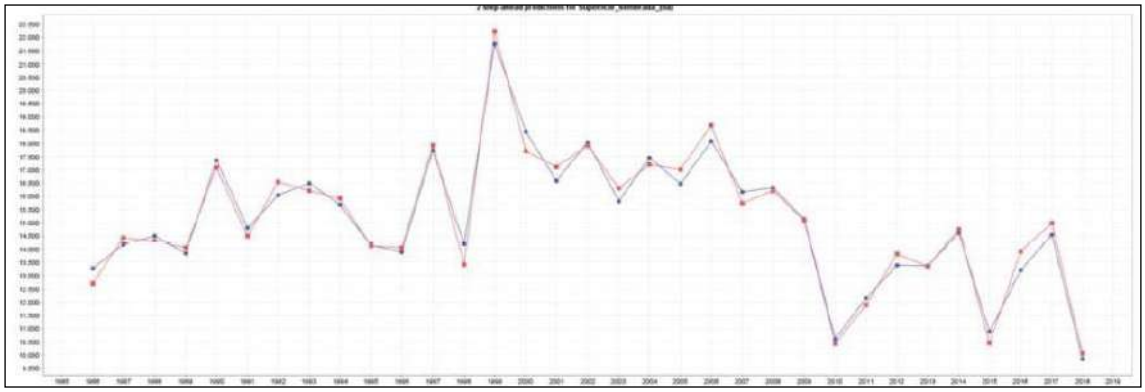
Los resultados se muestran a continuación y son parte de una red neuronal, misma que se visualiza gráficamente para el usuario final.

Imagen 10: Red neuronal para la producción de trigo en Biobío



Fuente: Zamora, 2018

Imagen 11: Red neuronal para la siembra de papa en Los Ríos y Los Lagos



Fuente: Zamora, 2018

Finalmente, el autor concluye en que los modelos que se han aplicado a los datos seleccionados no son los adecuados, sin embargo, al hacer una comparación entre los distintos modelos se obtiene que las redes neuronales es el mejor modelo debido a que se están trabajando con datos no lineales y eso puede influir en el comportamiento de estos modelos.

2.2. MARCO TEÓRICO

2.2.1. MINERÍA DE DATOS

(Suarez et al., 2022) plantea lo siguiente: “La minería de datos es una técnica que usa inteligencia artificial, aprendizaje automático, estadísticas y sistemas de bases de datos para encontrar patrones en grandes cantidades de datos (…)”.

Se puede decir que la minería de datos emplea entre otras cosas la estadística como base de su uso en determinadas circunstancias. Sin embargo, esto no quita que este estudio siempre esté presente cuando se refiere al uso de la virtualización (visualización gráfica) o la predicción (inteligencia artificial).

En la actualidad la mayor parte de MYPES hacen uso de distintos sistemas tecnológicos para estar al tanto del trabajo continuo que realizan, esto significa el recojo y almacenamiento de una gran cantidad de datos que incluyen perfiles de usuarios, clientes, transacciones, actividades, descripciones, etc. Los cuales resultan masivos en algún momento ya que esta información va en aumento.

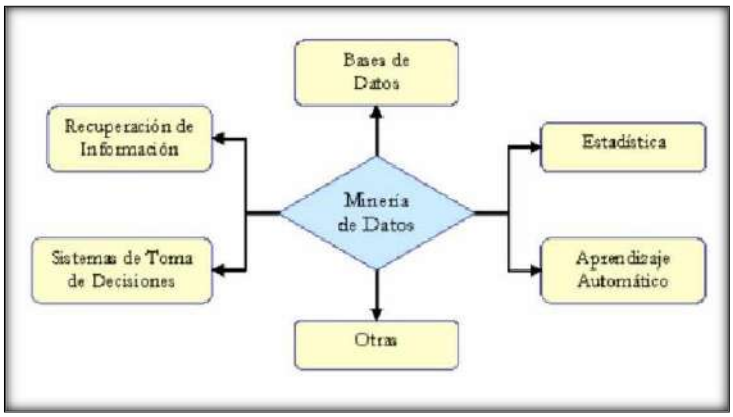
Los sistemas tecnológicos o base de datos se dan por la necesidad de salvaguardar información de la gran cantidad de datos que se posee, para en algún momento analizarlos y poder tomar una decisión mucho más eficiente y rápida.

Los datos almacenados no hacen más que describir estos mismos de manera cualitativa o cuantitativa y que puede ser usada en la estadística descriptiva para una visualización de tablas o gráficos que simplemente organiza y presenta un conjunto de datos. Sin embargo, esta información no es del todo relevante en las organizaciones que desean sobresalir y enmendar problemas que no están a simple vista.

Estos datos comienzan a tener valor útil al momento de ser analizados a profundidad con ayuda de la computación, ya que manualmente sería imposible determinar algún patrón. El proceso de llevar los datos a información relevante y conocimiento importante para tomar buenas decisiones toma el nombre de Minería de Datos o Data Mining, considerado como campo de la estadística y ciencias de la computación. Se aprecia en la

Imagen 1 la división para la Minería de datos en cuanto a su definición y comprensión.

Imagen 12: Definición de minería de datos.



Fuente: Galán, 2019

La minería de datos se considera un enfoque por el cual, de forma sincronizada, se puede encontrar una solución viable que será beneficiosa para aumentar el crecimiento. Es así como los agricultores del sector agrícola se enfrentan muchos problemas y obstáculos que se presentan a la inadecuada comprensión y aplicación de las actividades para mejorar la productividad.

De acuerdo con (Yash V. Bagal, 2020): “Las empresas agrícolas tienen la capacidad de recopilar y generar una gran cantidad de datos, de los que se extraen los requeridos mediante la automatización”. Es entonces que podemos saber que a partir de este punto entra en acción la minería de datos, que se puede usar para estudiar y predecir rasgos futuros en todos los aspectos del proceso de agricultura.

Los métodos de extracción de datos más usados en la minería de datos basándonos en los datos y tipo de datos que se pueden recoger de los procesos en el sector agrícola, son:

- Agrupación.
- Reglas de asociación.
- Clasificación.
- Tamaño de mercado.
- Regresión.

(Bodero-Poveda et al., 2022) afirma que: “(…) La minería de datos es un campo de investigación multidisciplinario interdisciplinario orientado a aplicaciones, combinando teorías y tecnologías en campos diferentes, como el aprendizaje automático, las bases de datos y las estadísticas matemáticas

(…)". Dicho autor hace mención de algunos algoritmos de minería de datos los cual considera son los más utilizados.

Métodos	Aspectos importantes	Ventajas	Desventajas
Redes neuronales	- Operan de forma paralela y permite desarrollar tareas cognitivas (aprendizaje de patrones, predicciones, clasificación y optimización) - Tipo de aprendizaje no supervisado.	- Descubrimiento y extracción de conocimiento - Clasifican conjuntos de datos. - Tolerante a fallas. - Márgenes bajos de error. - Clasifican patrones desconocidos con patrones conocidos que comparten las mimas características distintivas. - Funciona a pesar de contener fallas significativas en su estructura.	- Complejo para aprendizaje de tareas grandes. - No interpreta resultados. - Puede presentar extensos tiempos para el aprendizaje.
Análisis Clúster	- Construcción y clasificación de un conjunto de datos en grupos similares - Tipo de aprendizaje no supervisado	- Agrupa datos similares. - Reduce complejas cantidades de información en grupos pequeños con características similares. - Fácil de entender, aplicar y adaptar. - Eficiente en cuanto desempeño y resultados. - Se aplica a datos estáticos	- Algunas técnicas de clústering no son convenientes para grandes grupos de datos y se necesita que el usuario defina el número de clustering. - Puede llegar a ser computacionalmente costoso.

		y datos dinámicos.	
Árbol de decisión	- Análisis mediante un esquema facilitando la toma de decisiones en los resultados y probabilidades asociadas. - Tipo de aprendizaje supervisado.	- Fáciles de representar e interpretar. - Permite llegar a la toma de decisiones o acciones específicas. - Resultados más precisos. - Reduce el número de variables independientes.	- Seleccionar el atributo adecuada para la raíz del árbol. - No es práctico utilizar para pequeños conjuntos de datos. - Propenso a sobreajuste u overfitting. (demasiada información utilizada en el modelo)
Modelos de regresión	- Tenemos entre los modelos de regresión: Lineal, múltiple, hedónica, ponderada geográfica, ponderada temporalmente, crestas, polinomial. - Tipo de aprendizaje supervisado.	- Existen diversas técnicas aplicadas a distintos procesos u objetivos. - Fácil de desarrollar y entender. - Útil para comprensión de datos y toma de decisiones. - Útiles para la clasificación de texto e imágenes.	- Puede presentar valores atípicos. - Deficiente para realizar modelaciones complejas.
Bosque aleatorio	- Compuesto por múltiples arboles de decisión para obtener predicciones más precisas y estables. Reduce las altas varianzas de árboles individuales. - Tipo de aprendizaje supervisado.	- Simple de utilizar. - Procesamiento de variables cualitativas y cuantitativas sin necesidad de variables indicadoras. - Valida información al mismo tiempo que construye árboles, generando	- Ausencia de interpretabilidad. - Propenso a overfitting u sobreajuste. (demasiada información utilizada en el modelo) - Su proceso puede ser lento por presentar numerosos árboles de decisión.

		estimaciones de errores.	

Tabla 9: Técnicas más usadas de minería de datos

Fuente: Bodero-Poveda et al., 2022

2.2.2. APRENDIZAJE AUTOMÁTICO

La inteligencia artificial ha crecido de forma exponencial en los últimos años y esto de la mano del aprendizaje automático, nos referimos en este punto al contexto del análisis de datos y de la computación en general, lo que a su vez permite que las aplicaciones funcionen de manera inteligente.

En este punto cabe preguntarse qué es o a que hace referencia el aprendizaje automático o Machine Learning por su traducción al inglés. En palabras de (Sarker, 2021): "(…) Machine Learning brinda a los sistemas la capacidad de aprender y mejorar a partir de la experiencia automáticamente sin ser programados específicamente y generalmente se lo conoce como el más popular de las últimas tecnologías en la denominada 4ta revolución industrial o industria 4.0.".

Podemos entender entonces que para "ingresar" o mantenerse en la nueva era de la información y de la industria como la conocemos hoy en día tenemos que estar de la mano del aprendizaje automático, es decir, para analizar de forma inteligente los datos y desarrollar aplicaciones que funcionen en el día a día, los algoritmos y todo el ecosistema que engloba a Machine Learning serán la pieza fundamental.

De manera breve y antes de pasar a desarrollar la teoría de forma más amplia, veamos que piensan otros autores sobre lo qué es aprendizaje automático, tal como lo describe (Ccoyccosi, et al., 2021): "(…) ML es la ciencia que se hace a partir de datos para que los ordenadores aprendan. En la programación convencional se programa paso a paso cada solución en particular para cada necesidad que fue planteada, el área de ML se dedica a desarrollar algoritmos genéricos los cuales pueden obtener

patrones de varios tipos de datos (…)". Lo que este autor también nos hace hincapié en mencionar que para el proceso de aprendizaje automático o automatizado se siguen distintas etapas, mismas que se ilustran en el siguiente gráfico:

Imagen 13: Etapas de construir un modelo de Machine Learning
Fuente: Manrique, E. (2020).

Como hemos hablado hasta el momento, el proceso de construir un modelo de Machine Learning requiere de datos y generalmente estos se consumen, así como se procesan, de tal forma se aprenden y encuentran los patrones relacionados a las personas, procesos, transacciones, etc. Es entonces que tenemos que considerar ¿Qué datos requiere un modelo de aprendizaje automático?

Hay que considerar que los datos son de distintos tipos que se dividen en estructurados, semiestructurados o no estructurados, de la tal forma podemos observar la diferencia, definición y características en la siguiente tabla:

Tipos de datos		
Datos estructurados	**Datos semi estructurados**	**Datos no estructurados**
De acuerdo con (Sarker, 2021): "Los datos estructurados tienen una estructurada bien definida, se ajusta a un modelo de datos siguiente un estándar	En palabras de (Sarker, 2021):"Los datos semiestructurados no se almacenan en una BD relacional, pero tiene ciertos propiedades	Como expresa (Sarker, 2021): "No existe un formato predefinido para los datos no estructurados, lo que nos hace mucho más difíciles de capturar, procesas y analizar, ya

en esquemas bien definidos, como BD's relacionales".	organizativas que la hacen más fácil de analiza, algunos ejemplos son HTML, XML, JSON, BD's NoSQL, etc".	que en su mayoría contienen texto y material multimedia".
Por tanto, alguna de las características serían las siguientes: - Se ingresan siguiendo una estructura. - Cada columna tiene distintos tipos de valores. - Alrededor del 20% de datos en el mundo están estructurados.	Por tanto, alguna de las características serían las siguientes: - Se asemejan a los datos estructurados. - Es más flexible, pero siguen siendo estructurados. - Otro formato de almacenamiento es YAML.	Por tanto, alguna de las características sería de las siguientes: - No siguen un modelo y no están contenidos en filas y columnas. - Difícil de buscar y organizar. - Mayormente son texto, sonido, videos o imágenes. - Estos datos pueden ser muy valiosos.

Tabla 10: Tipos de datos del mundo real

Fuente: Elaboración propia

Bien, una vez hemos visto los tipos de datos lo que se hace con esto es la construcción de un modelo y para ello se usan distintas técnicas y estas se dividen generalmente en cuatro categorías, mismas que son las siguientes: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje por refuerzo. Veremos una visión más amplia en la siguiente tabla:

Tipos de técnicas de aprendizaje automático		
Aprendizaje	**Descripción**	**Ejemplos**
Supervisado	Aprenden de los datos etiquetados (basado en tareas).	Clasificación, regresión.
No supervisado	Aprenden de datos no etiquetados (basado en datos).	Clustering, asociaciones, reducción de dimensionalidad.
Semi supervisado	Se construyen usando datos combinados (etiquetados y no etiquetados).	Clasificación, agrupamiento.

Reforzamiento	Se basan en la recompensa o la penalización (basado en el entorno).	Clasificación, control.

Tabla 11: Tipos de técnicas de Machine Learning

Fuente: Elaboración propia

(Chhaya et al., 2020) considera que: “El aprendizaje automático además de resolver problemas también beneficia a las organizaciones al hacer predicciones para ayudarlos a mejorar la toma de decisiones, gracias al aprendizaje automático se podrá saber notoriamente dónde utilizarlo y dónde no.” Es necesario identificar las ventajas y desventajas de los lenguajes del aprendizaje automático, estos se muestran a continuación”.

Ventajas	Identifica fácilmente tendencias y patrones	Administra una gran cantidad de datos y comprende las tendencias y patrones que no sería posible ser analizado por los humanos.
	No necesita interferencia humana	No es necesario ayudar al sistema o darle comandos para seguir instrucciones.
	Mejora continua	El algoritmo del aprendizaje automático comprende continuamente los errores y los rectifica. Aumenta la eficiencia y precisión.
	Manejo de datos multidimensionales y de múltiples variedades	Administra y mejora la gran cantidad de datos multidimensionales y mejora sus habilidades para evitar errores.
	Amplias aplicaciones	El uso del aprendizaje automático puede ser útil para distintos campos de comercio electrónico.
	Adquisición de datos	La gran cantidad de datos que se usen en el proceso de formación y aprendizaje debe de ser de buena calidad.

Desventajas	Tiempo y recursos	Durante el procedimiento de aprendizaje automático, los algoritmos para administrar los datos pasan por un proceso de rectificación de errores, y ello requiere tiempo. Por otro lado, los recursos deben ser confiables para el funcionamiento.
	Interpretación de resultados	Se debe verificar el resultado y seguir la operación de corrección para obtener la información deseada.
	Alta susceptibilidad a errores	En el proceso de aprendizaje automático se utilizan y prueban muchos algoritmos. Cuando ocurren errores no es fácil descubrir la fuente principal por la se creó el problema.

Tabla 12: Ventajas y desventajas del aprendizaje automático

Fuente: Chhaya et al., 2020

2.2.3. TÉCNICAS DE MINERÍA DE DATOS

(Alyahyan & Düştegör, 2020) argumenta: "(…) en aplicaciones como esta – se refiere a su trabajo de investigación – se usan mayormente los modelos de minería de datos más comunes, mismos que son dos y sirven para la predicción del éxito: predictivos y descriptivos (…)".

Podemos entonces describir que las técnicas más usadas son las mencionadas al principio de la cita, las predictivas y las descriptivas. Si nos basamos en la primera técnica podemos decir que son aquellas que tienen variable dependiente e independiente, puesto que se usan para análisis de dependencia o explicativos.

Ahora, las técnicas descriptivas serán aquellas que tienen el mismo estado, similar a las técnicas que analizan métodos descriptivos de análisis con muchas variables. Esto último lo explicaremos a detalle a continuación junto con las técnicas específicas de esta división en general de predictivas y descriptivas.

1. TÉCNICAS PREDICTIVAS

(Alyahyan & Düştegör, 2020) asegura sobre las técnicas predictivas: "(…) aplican funciones de aprendizaje supervisado para proporcionar una estimación de los valores esperados de las variables dependientes en función de las características de las variables independientes relevantes (…)".

Es decir, que tal como habíamos especificado líneas antes, esto se basa en una manera de inteligencia artificial, pero semi automática, puesto que ellas no aprenden por sí solas a medida que pase el tiempo si no tienen conocimientos previos.

Estos modelos definidos por las técnicas predictivas necesitan de información que proviene de algún lado para poder ser procesada y en definitiva sacarle provecho para obtener conocimiento concreto.

Es un hecho entonces que, al finalizar con el proceso de usar una técnica descriptiva, que la conclusión será contrastar la información que se tenía antes de usar la predicción como técnica y diferenciarla del conocimiento tras su uso.

Alguna de las técnicas predictivas más usadas son las mencionadas a continuación.

ÁRBOLES DE DECISIÓN

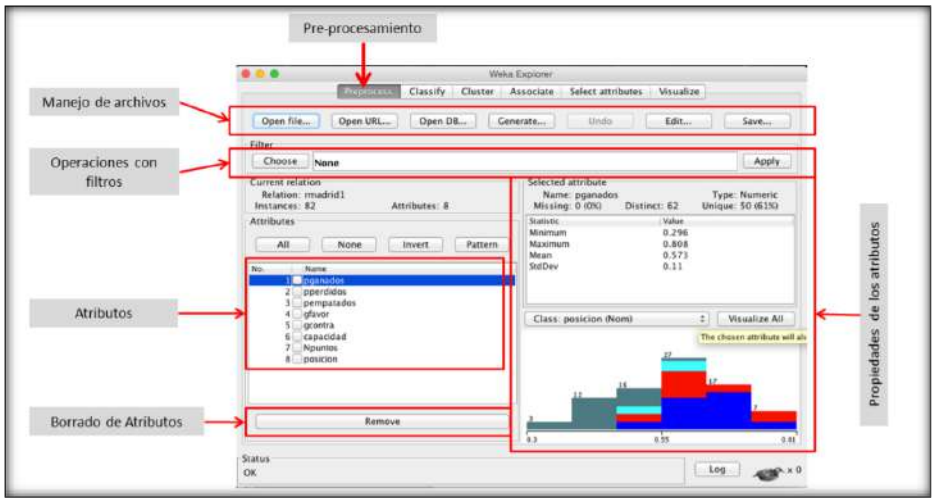
Es un algoritmo que categoriza la información con objeto de generar un modelo en forma de árbol, así esquematizando información de distintas alternativas junto a posibles resultados.

- **En Weka:**

  Los pasos a realizar según el estudio de (Navas, 2016) son los siguientes:

  Paso 1. "Preprocesamiento de datos. Utilizando la página principal del Explorer en la pestaña de preprocess, es posible determinar filtros, atributos o instancias. Es posible realizar transformaciones a nuestros datos con el fin de realizar muestreos, unificar valores, normalizar valores numéricos, etc.". (Navas, 2016).

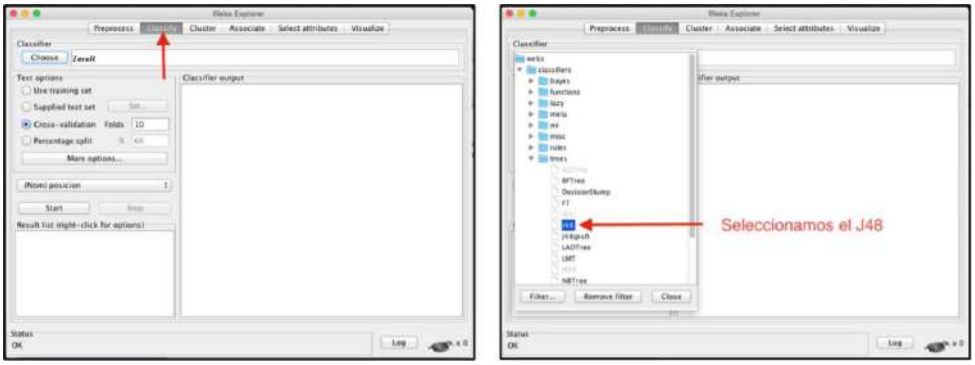
Imagen 14: Página principal de Explorer (WEKA)



Fuente: Navas, 2016

Paso 2. "Clasificación. Siendo el árbol de decisión un algoritmo de clasificación nuestro objetivo es encontrar relaciones entre los datos para conocer las posibilidades, este proceso en WEKA se lleva a cabo en la pestaña classify". (Navas, 2016).

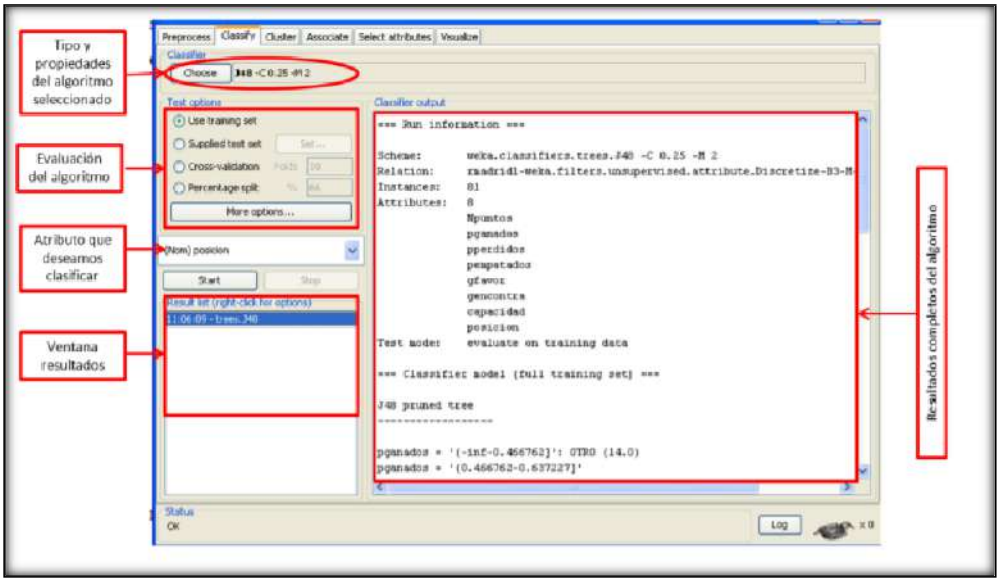
Imagen 15: Pestaña de classify y selección del árbol de decisión J48



Fuente: Navas, 2016

Al seleccionar el clasificador del árbol de decisión J48 se obtendrán la siguiente interfaz. En el cual para las opciones del test se tomará en cuenta la técnica de cross-validation, obteniendo con ello análisis estadísticos independientes.

Imagen 16: Árbol de decisión J48



Fuente: Navas, 2016

Paso 3. "Interpretación de resultados. Al completar el entrenamiento se obtendrá un resumen del test, dónde destacan una serie de valores estadísticos y la precisión obtenida". (Navas, 2016).

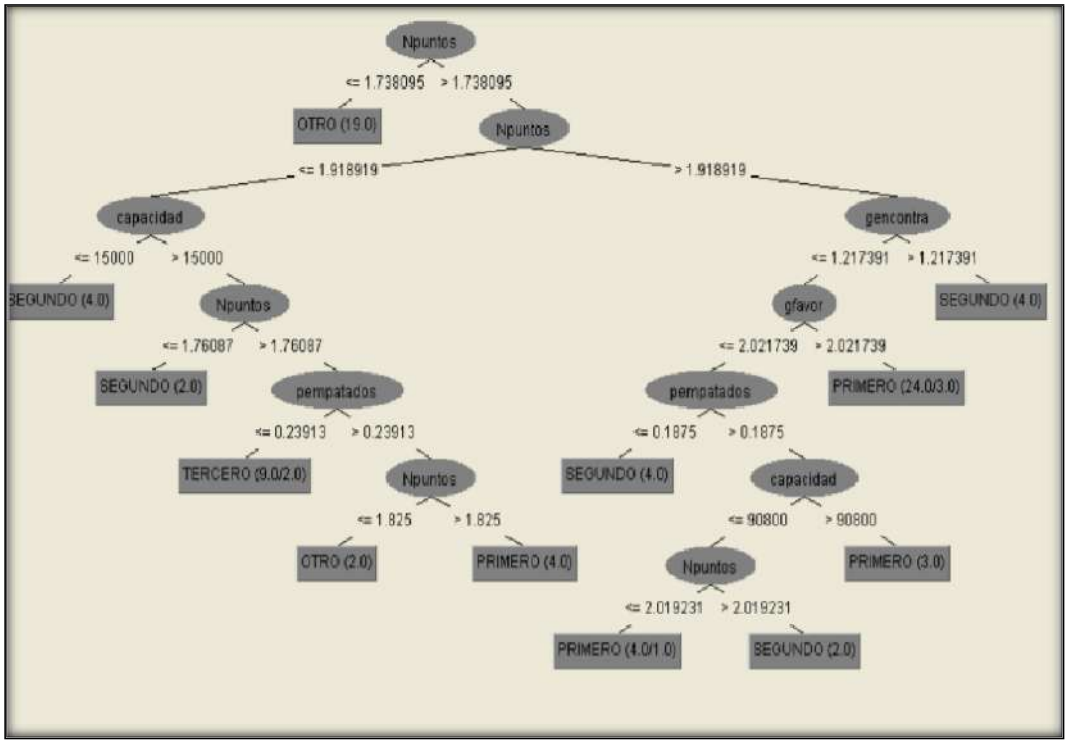
Imagen 17: Resultados del ejemplo, árbol de decisión J48

```
=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances          75               92.5926 %
Incorrectly Classified Instances         6                7.4074 %
Kappa statistic                          0.8943
Mean absolute error                      0.0622
Root mean squared error                  0.1764
Relative absolute error                 17.6828 %
Root relative squared error             42.1157 %
```

Fuente: Navas, 2016

Paso 4. "Visualización gráfica. Para una mejor vista de los resultados y mayor entendimiento del objetivo, es necesario acceder al gráfico y determinar conclusiones en base a nuestro dataset". (Navas, 2016).

Imagen 18: Resultado gráfico árbol de decisión J48.



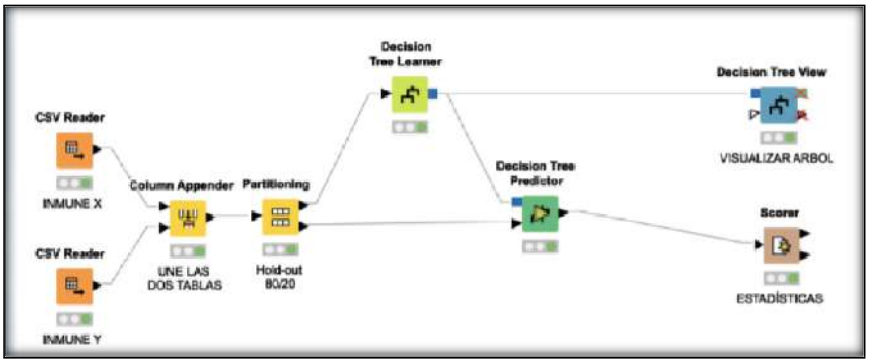
Fuente: Navas, 2016

- **En Knime:**

Los pasos a realizar según el estudio de (Soto & Martínez, 2020) son los siguientes:

Paso 1. "Selección de nodos. Se usarán los nodos Decision Tree Learner (aprender) y Decision Tree Predictor (predecir), además de leer los datos con un CSV Reader y Partitioning para configurar que porcentaje queremos entrenar y lo demás para la prueba. Por último, el Scorer mostrará los datos estadísticos". (Soto & Martínez, 2020).

Imagen 19: Flujo de datos para árbol de decisión en KNIME.



Fuente: Soto & Martínez, 2020

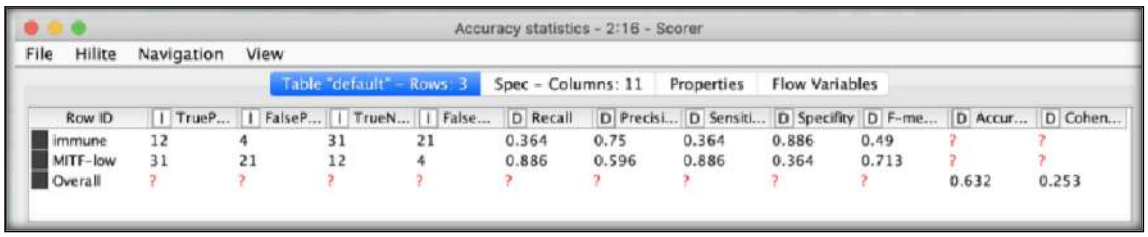
Paso 2. “Configuración de los nodos. En cuanto al nodo Partitioning este se lleva a cabo en un 70% para el entrenamiento y 30% para la prueba. En cuanto al aprendizaje, La siguiente imagen muestra las opciones de parámetro de configuración del nodo Decision Tree Learne”. (Soto & Martínez, 2020).

Imagen 20: Configuración del nodo Decision Tree Learner

Fuente: Soto & Martínez, 2020

Paso 3. “Resultados. Los resultados a mostrar con el nodo Scorer son las estadísticas de precisión dónde se mostrará como en la siguiente imagen.”. (Soto & Martínez, 2020).

Imagen 21: Estadísticas de precisión (Scorer)

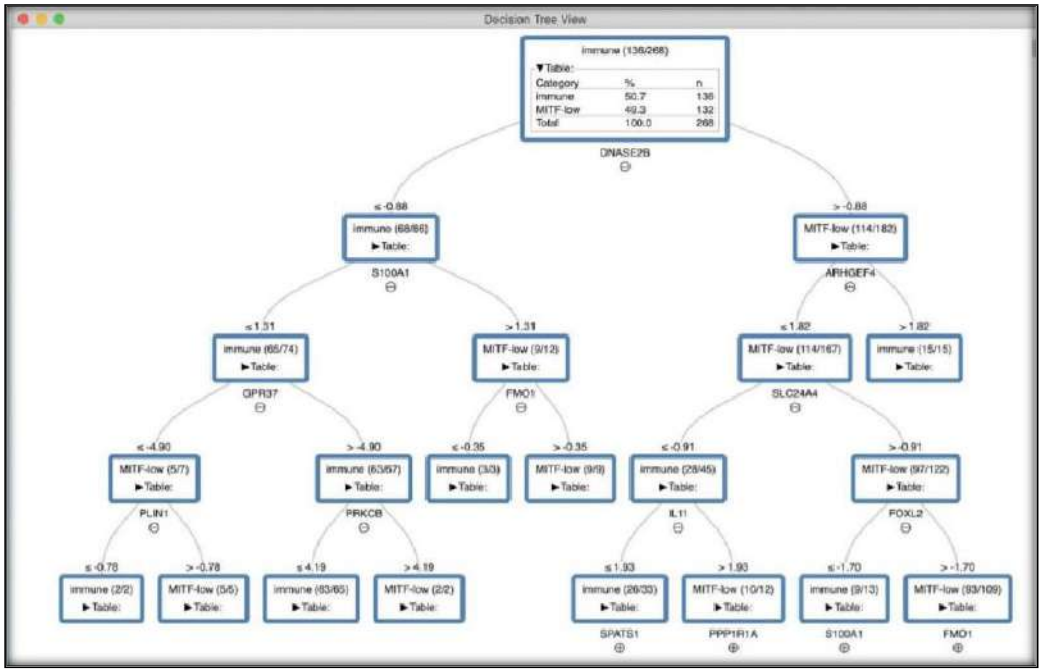


Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accur...	Cohen...
immune	12	4	31	21	0.364	0.75	0.364	0.886	0.49	?	?
MITF-low	31	21	12	4	0.886	0.596	0.886	0.364	0.713	?	?
Overall	?	?	?	?	?	?	?	?	?	0.632	0.253

Fuente: Soto & Martínez, 2020

Paso 4. “Visualización Gráfica. Después de llevar a cabo las configuraciones necesarias el árbol se mostrará como en la siguiente imagen, al dar clic derecho en el Decision Tree Predictor y en View: Decision Tree View”. (Soto & Martínez, 2020).

Imagen 22: Decision Tree View

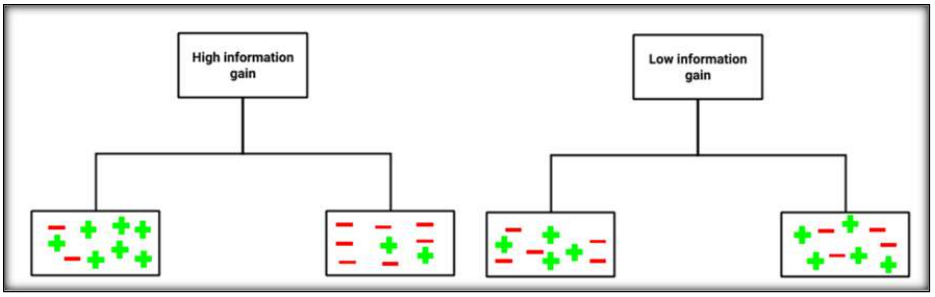


Fuente: Soto & Martínez, 2020

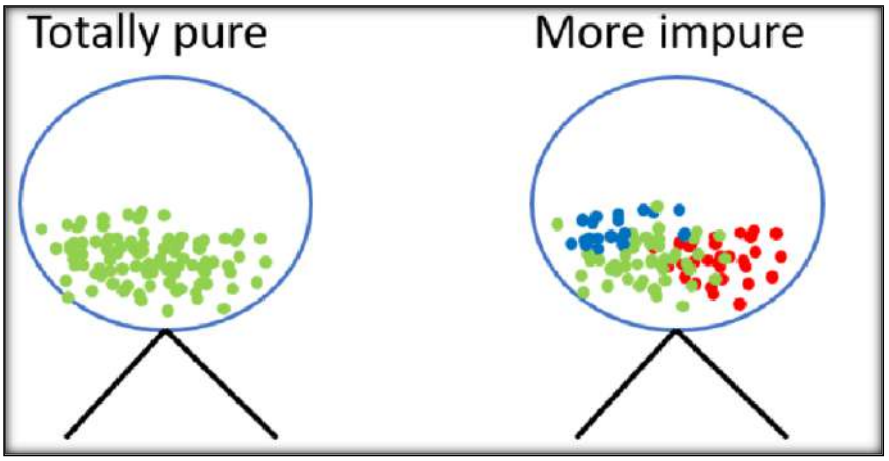
- **Matemática y estadística detrás de los árboles de decisión:**

Las medidas más importantes a tomar en cuenta según (Acevedo, 2020)a para llevar a cabo este punto vienen a ser ganancia de información y el índice GINI:

a. Ganancia de información. Es necesaria para medir que tan bien un atributo ha sido separado en el entrenamiento, siendo estos de tipo puro (requiere menos información, valores similares), impuro (requiere más información para ser descrito) y más impuro (requiere información máxima). Al tener un atributo con alta ganancia de información, los datos se dividirán en grupos desiguales de positivos y negativos, con más facilidad para separarlos entre sí. Mientras que una baja ganancia de información resultará en datos uniformes y que no nos acercarán a una fácil decisión.



Entropía. Utilizado para medir la impureza de los datos de entrada, la ganancia de información es una disminución de entropía, mientras menor sea el valor de la entropía, los datos serán más puros.



En conclusión, cuanto más impuro sea el conjunto de datos, mayor será la entropía y al ser menos impuro, menor será la entropía. Si la entropía es 0, todos los datos pertenecen a la misma clase, por lo que Entropía (S) = 0.

Fórmula para la entropía

$$Entropy\ (S) = (-p * log2\ x\ p) - (q\ x\ log2\ x\ q\ )$$

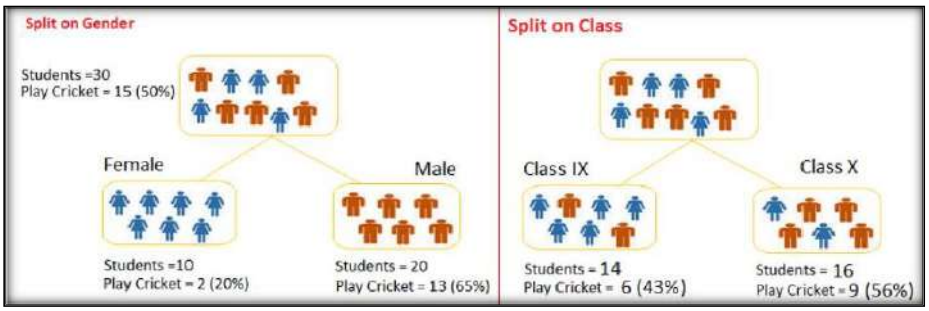
Donde:
P, es probabilidad de éxito.
Q, es probabilidad de fracaso.

Entropía de una división

Para calcular la entropía en una división es necesario calcular la entropía del nodo padre, la entropía de cada nodo individual de la división y calcular el promedio ponderado de los subnodos disponibles.

$$Información\ obtenida = Entropía\ (nodo\ principal) - [Promedio\ de\ entropía\ (hijos)]$$

Ejemplo:
Muestra = 30 estudiantes
Variables = 3 – (Sexo (Niño / Niña), Clase (IX / X) y Altura (5 a 6 pies)
15 de estos 30 juegan al cricket en el tiempo libre.



$$Entropía\ para\ el\ nodo\ principal$$
$$= -\ (15/30)\ log2\ (15/30) - (15\ /\ 30)\ log2\ (15/30)$$
$$= 1$$

Aquí el valor uno (1) muestra que es un **nodo impuro.**

Entropía para género:
Entropía para nodo femenino = – (2/10) log2 (2/10) – (8/10) log2 (8/10) = 0,72 Entropía para el nodo masculino = – (13/20) log2 (13/20) – (7/20) log2 (7/20) = 0.93
Entropía por división Género = (10/30) * 0.72 + (20/30) * 0.93 = 0.86

Resultado Ganancia de Información:
Ganancia de información por división por género = 1–0.86 = 0.14

Entropía para Clase:
Entropía para nodo de clase IX = – (6/14) log2 (6/14) – (8/14) log2 (8/14) = 0.99 Entropía para clase Nodo X = – (9/16) log2 (9/16) – (7/16) log2 (7/16) = 0.99 Entropía para división Clase = (14/30) * 0.99 + (16/30) * 0.99 = 0,99

Resultado Ganancia de Información:
Ganancia de información por división por clase = 1– 0.99 = 0.01

Conclusión: La ganancia de información para género es más alta, por lo tanto, el árbol se dividirá en género.

Ganancia de información:
Utilizado para minimizar la profundidad de un árbol de decisión, seleccionando un atributo óptimo para dividir el árbol.
Este atributo seleccionado debe tener una mayor reducción de la entropía.
La ganancia de información es la reducción de la entropía esperada relativa a un determinado atributo.

GI Fórmula:
Ganancia de información, Gain (S, A) de un atributo A.

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v=1}^{v=N} \frac{|\ Sv\ |}{|\ S\ |}\ x\ Entropy\ (Sv)$$

Algoritmo C4.5 (J48):

La medida GI favorece aquellas variables con mayor número de posibles valores. Dado S (conjunto de ejemplos de aprendizaje) y A (atributo de los ejemplos con c valores) se define:

$$SplitInformation(S, A) = -\sum_{i=1}^{c} \frac{|\ Si\ |}{|\ S\ |} \log 2\ \frac{|\ Si\ |}{|\ S\ |}$$

SplitInformation (S, A) denota la entropía S con respecto a valores de A.

$$RatiodeGanancia(S, A) = \frac{Ganancia\ (S, A)}{SplitInformation(S, A)}$$

RatiodeGanancia (S, A) favorece aquellos atributos para ser separados en menos clases.

b. GINI.

La población según GINI será pura, siempre y cuando al seleccionar dos elementos al azar de una población, estos deben ser de la misma clase y la probabilidad sea 1.

- GINI funciona con variables objetivo categórica “Éxito” o “Fracaso”.
- Realiza divisiones binarias.
- Mientras GINI sea mayor, mayor será la homogeneidad.

Calcular GINI para una división:

- Usar suma de la fórmula del cuadrado de probabilidad de éxito y fracaso

$$(p^2 + q^2)$$

- Usar puntuación ponderada de GINI para cada nodo.

Para el mismo ejemplo mencionado anteriormente. Ahora se requiere identificar la división que está produciendo subnodos más homogéneos usando GINI.

División en género:

Gini para el sub-nodo Mujer:

$$= (0.2) * (0.2) + (0.8) * (0.8) = 0.68$$

Gini para sub-nodo Hombre:

$$(0.65) * (0.65) + (0.35) * (0.35) = 0.55$$

Calcular Gini ponderado para Split Gender:

$$(10/30) * 0.68 + (20/30) * 0.55 = 0.59$$

División en clase:

Gini para sub-nodo Clase IX:

$$(0.43) * (0.43) + (0.57) * (0.57) = 0.51$$

Gini para sub-nodo Clase X:

$$(0.56) * (0.56) + (0.44) * (0.44) = 0.51$$

Calcular Gini ponderado para Split Class:

$$(14\ /30) * 0.51 + (16/30) * 0.51 = 0.51$$

Gini Split Gender > Gini Split Class

Por lo tanto, la división de nodos tendrá lugar en género.

Impureza de Gini

Se determina restando el valor de Gini en 1. En términos generales:

$$Impureza\ de\ Gini = 1 - Gini$$

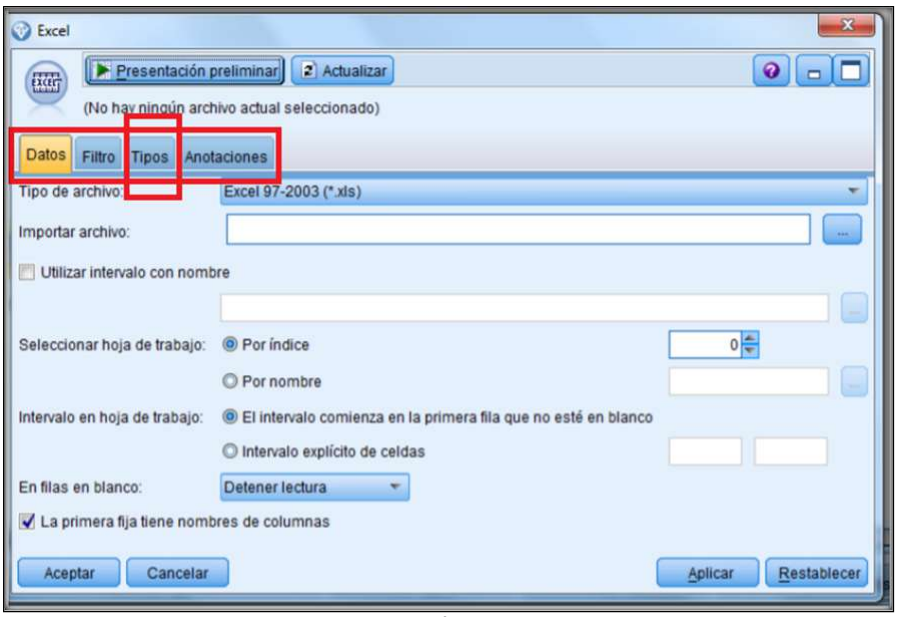
MÉTODOS BAYESIANOS

Estos métodos, también llamados estimaciones, tienen la capacidad se sintetizar, simplificar y resumir la información de muestras y la no muestral usando el teorema de Bayes.

- **En SPSS Modeler:**

Para la implementación de redes bayesianas se debe tener en cuenta que es un proceso que parte desde la lectura de los datos de estudio en la paleta de nodos de origen, el cual desplegará la interfaz mostrada en la siguiente imagen.

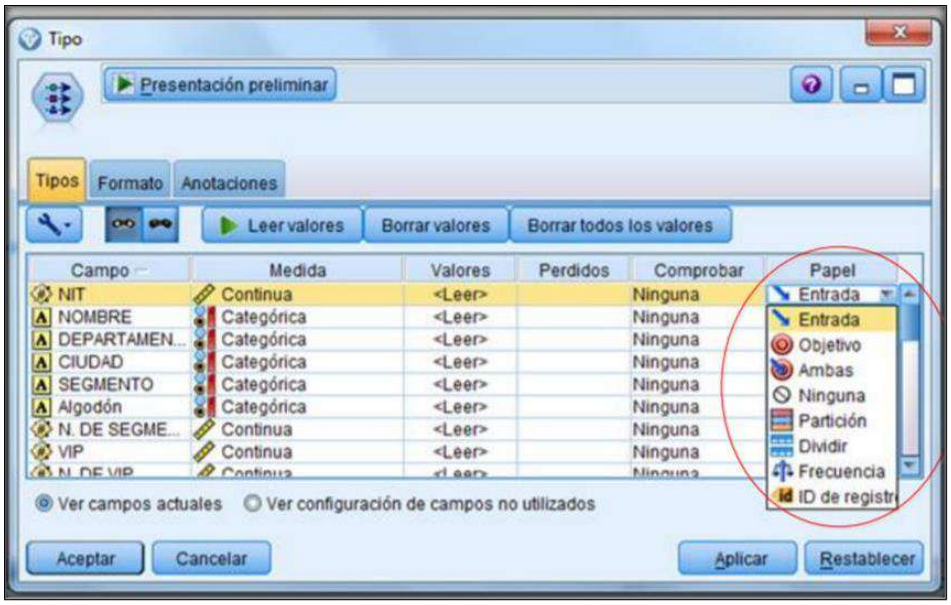
Imagen 23: Datos de entrada nodo origen (SPSS Modeler)



Fuente: Rivero, 2012

El siguiente paso es definir el tipo de las variables, en la misma interfaz al seleccionar la pestaña de tipos podremos personalizar y leer los valores. Es importante seleccionar correctamente el papel de cada uno de ellos.

Imagen 24: Personalización de nodo tipo (SPSS Modeler)



Fuente: Rivero, 2012

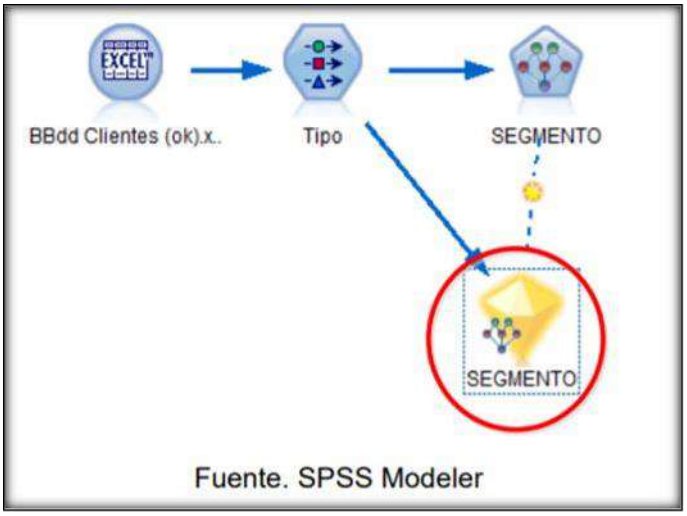
A continuación, se selecciona el nodo de red bayesiana para llevar a cabo el modelado.

Fuente: Rivero, 2012

Al conectarla con el nodo anterior, obtendremos los parámetros a configurar de la red bayesiana, el cual es la siguiente:

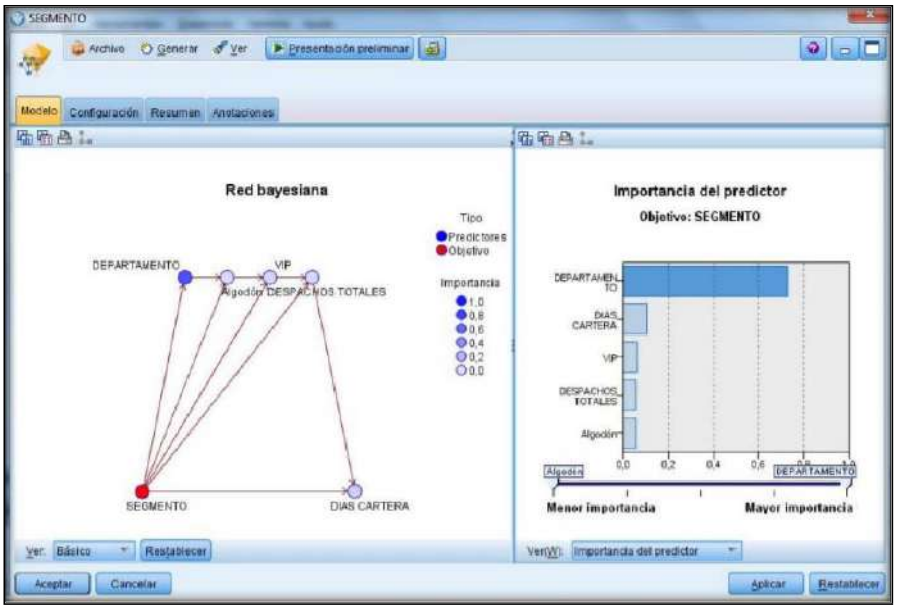
Fuente: Rivero, 2012

Al ejecutar el modelo se despliega un nuevo cuadro que crea la red bayesiana:



Fuente: Rivero, 2012

Al hacer clic sobre el nuevo segmento creado, se podrá visualizar las probabilidades como resultado. Los resultados son interpretados de acuerdo con los datos cargados.



Fuente: Rivero, 2012

- **Matemática y estadística detrás de las redes bayesianas:**

La estadística Bayesiana según (Mesa et al., 2018) estudia mediante un análisis de datos la mejor manera de actuar, incorporando conocimiento de eventos pasados para poder obtener parámetros que provengan de la experiencia.

De qué manera se explicaría entonces el teorema de bayes, se resuelve en las siguientes líneas:

Siendo la fórmula básica para la probabilidad condicional la siguiente:

$$P(A|B) = \frac{P(B \cap A)}{P(A)}$$

Este método es utilizado para obtener diversos resultados de probabilidad condicional.

a. <u>Conceptos previos del teorema de bayes:</u>

- Sucesos o Eventos. Subconjunto del espacio muestral: Ω
- Sucesos mutuamente excluyentes. Dos sucesos son mutuamente excluyentes si estos ocurren a la vez: A ∩ B = Ø
- Las probabilidades están determinadas de acuerdo al carácter de dicha distribución siendo estas:

  i. Sea A un evento nulo: $A = \emptyset\,, P(A) = P(\emptyset) = 0$
  ii. Sea $\{A1, A2, \ldots, Ai, \ldots, An\}$ sucesos mutuamente excluyentes.

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} p(A_i)$$

iii. Sean A y Ac los eventos complementarios esto es A U Ac = Ω, es decir la unión de dos eventos: Para todo evento o suceso A en Ω

$$P(A) + P(Ac) = 1 \; o \; bien \; P(Ac) = 1 - P(A)$$

iv. Sea (A U B) el evento definido como que ocurre A o bien ocurre B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

b. <u>Demostración del Teorema de Bayes:</u>

Probabilidad de alguno de los eventos de A dado el caso el evento B se denota: $P\ (Ai,/\ B)$.

La probabilidad condicional es la siguiente:

$$P\left(\frac{A_i}{B}\right) . = P\ (B \cap \frac{(A_i)}{P(B)})$$

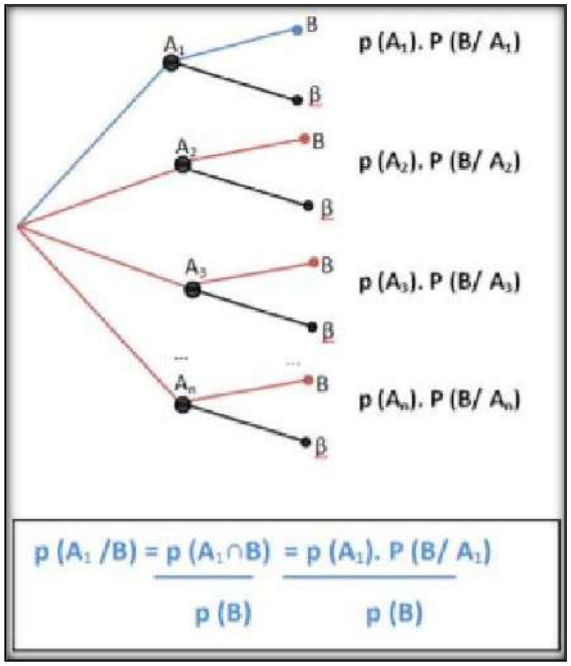
$$P\left(\frac{A_i}{B}\right) . = P_{A_i}^{(\ )}\ P\ (\frac{B}{A_i})\ /P(B)$$

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P(\frac{B}{A_i})}{\sum_{i=1}^{n} P(A_i)P(\frac{B}{A_i})}$$

c. <u>Ejemplo del Teorema de Bayes:</u>

Concordancia entre probabilidades de que algún suceso ocurra a partir de un suceso producido.

Imagen 25: Ejemplo teorema de bayes



Fuente: Rivero, 2012

2. TÉCNICAS DESCRIPTIVAS

(Alyahyan & Düştegör, 2020) manifiesta lo siguiente sobre las técnicas descriptivas: "(…) se usan para producir patrones que describen la estructura fundamental, las relaciones y la interconexión de los datos extraídos mediante la aplicación de funciones de aprendizaje no supervisado (…)".

Dando alusión a la cita anterior, entonces debemos recurrir a interpretaciones tales como la creación manual o automática con base en los patrones.

Sin embargo, lo descrito anteriormente no quita el objetivo ni diferencia entre esta técnica y otra el enfoque directo en descubrir conocimiento que esté dentro u oculto en los datos.

Una de las características más comunes de todo enfoque y/o tendencia descriptiva son que definen la clasificación, relación y delimitaciones, si las hubiera, entre las variables en una situación determinada.

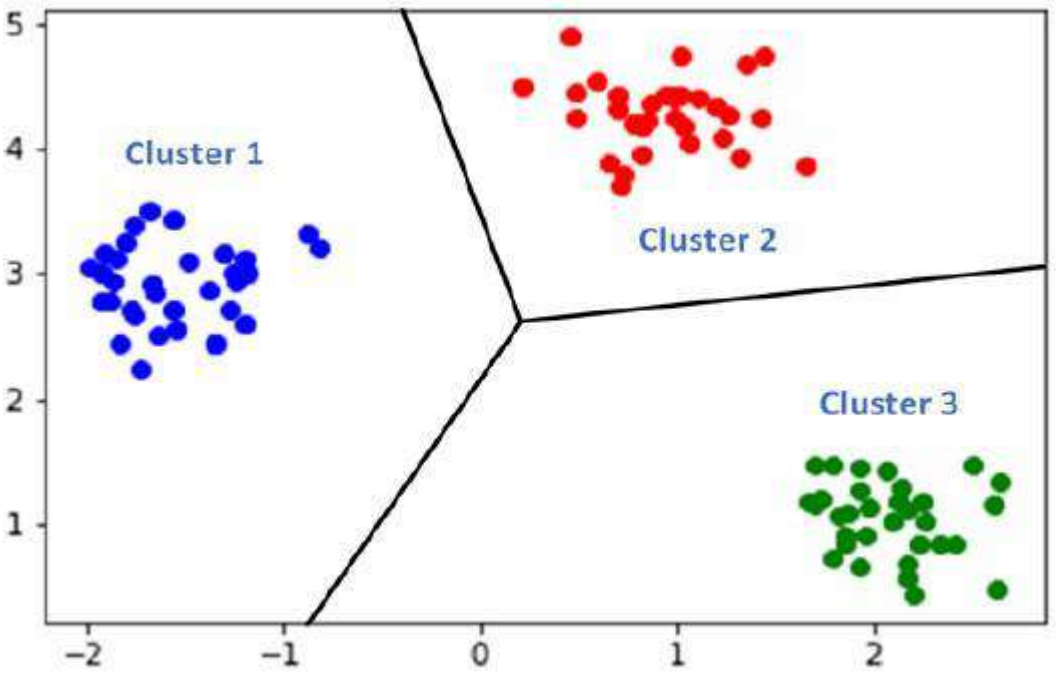
A continuación, se plasmará las técnicas descriptivas más comunes.

CLUSTERING

Puede ser vista como herramienta independiente para obtener ideas sobre distribución de datos. Al aplicarlo, de forma correcta, nos

proporciona agrupamientos en conjunto de datos y así poder observar colecciones de datos.

Imagen 26: Clasificación de datos usando la técnica cluster



Fuente: Ratz, A., 2020

DEPENDENCIA

Son patrones que establecen uno o más atributos que determinan explicaciones causales a partir de un conjunto de datos, con fin de usar dichas explicaciones para la descripción.

ANÁLISIS EXPLORATORIO

Como tal es un proceso por el cual el número de variables aleatorias se ve reducida del conjunto de datos que tiene como categorización una baja consideración. Se realiza mediante la obtención de un conjunto de variables principales.

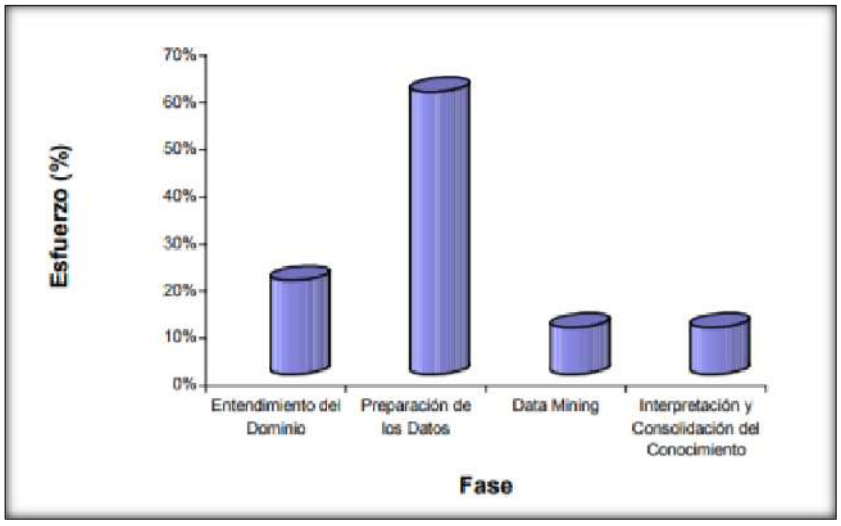
2.2.4. PROCESO KDD

Según (Edastama et al., 2021) la minería de datos es el proceso de extraer información y experiencia a partir de grandes cantidades de datos. La minería de datos es un conjunto de procedimientos para determinar el valor añadido de un conjunto de datos en forma de información no conocida previamente a partir de los datos. En el mundo de las bases de datos, la minería de datos también se conoce como descubrimiento de conocimiento (KDD). KDD es un proceso que implica la recopilación, el análisis y la interpretación

La minería de datos es solo el comienzo del proceso extracción de conocimiento a partir de datos, también conocido como KDD (Knowledge Discovery in Databases), este proceso se basa en identificar los distintos patrones desconocidos de los datos que se volverán realmente útiles para la organización. (Papathanasiou et al., 2022) define: "(…) KDD está dividido en 5 pasos los cuales cada paso corresponde a un proceso diferente de captura y análisis de datos, englobando algoritmos de aprendizaje automático para procesar y extraer conocimiento de bases de datos, desde una selección (obtención de un conjunto o subconjunto de datos que formarán parte del análisis), pre procesamiento (comprende comprobar la calidad de los datos, limpieza de datos), transformación (normalización, agregación, creación de nuevos atributos, reducción de datos), Data Mining (obtener resultados deseados, descubrir nuevos patrones) e interpretación y evaluación (presentación de resultados de manera adecuada) (…)".

Como se puede observar en la imagen a continuación según (Molina et al., 2006), el esfuerzo mostrado por cada fase recae en la preparación de datos lo cual indica que es una fase crucial para tener éxito en el proceso.

Imagen 27: Esfuerzo requerido por cada fase del proceso KDD



Fuente: Molina et al., 2006

Entonces, con las citas anteriores podemos darnos cuenta que la minería de datos tiene un proceso el cual lo engloba con más actividades que se realizarán para lograr el objetivo, y este se llama KDD.

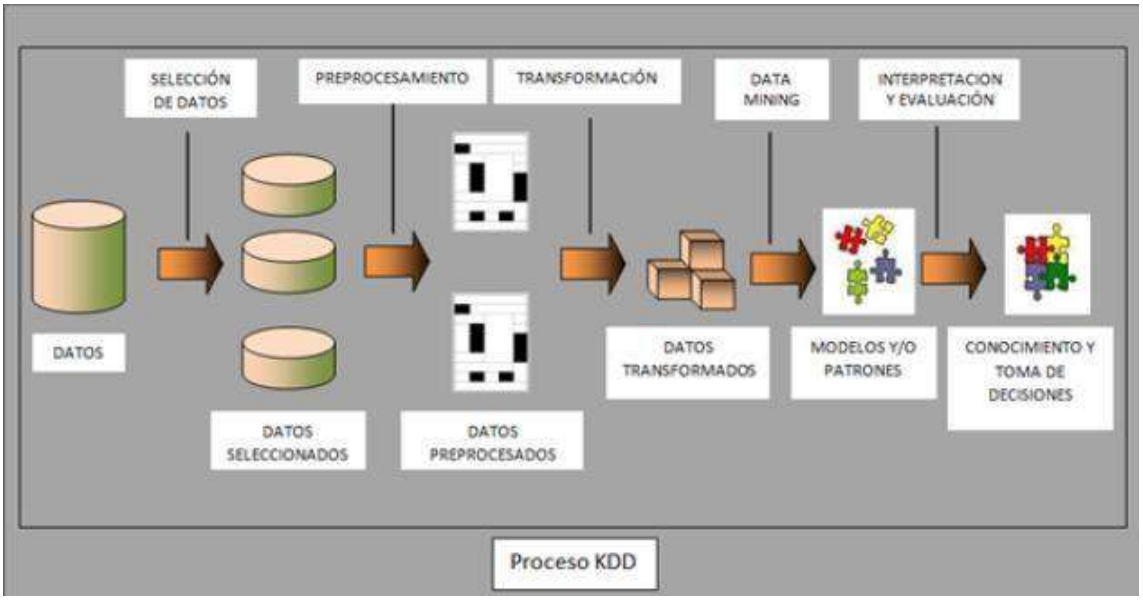
Generalmente el KDD tiene identificado el problema en principio para luego efectuar el preprocesamiento y en el camino se van definiendo las posibles actividades a realizar.

(Castro P. H., 2008) define al proceso KDD como: “Es un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos.”

Entonces, para definir claramente lo que es un KDD y que se traza como objetivos como poder tomar la cita anterior y reflexionar acerca de ello para culminar con que es un proceso para la obtención de conocimientos, sin embargo, tiene a la minería de datos como la pieza clave o parte principal de todas sus tareas.

Del mismo modo podemos identificar sus objetivos con claridad, los cuales vendrían hacer la generación de información y usar esta misma para darles beneficios competitivos a la empresa, se pueden dividir en varios: posicionamiento, utilidades, mejores estrategias, entre muchísimas otras.

Imagen 28: Proceso KDD



Fuente: Conde, D., 2017

Para conocer toda esta información de cada fase del proceso se ha resumido en la siguiente tabla.

Tabla 13: Resumen clasificación fases del proceso KDD

KDD							
SELECCIÓN	EXPLORACIÓN	LIMPIEZA	TRANSFORMACIÓN	MINERIA DE DATOS		EVALUACIÓN E INTERPRETACIÓN DE RESULTADOS	DIFUSIÓN Y USO DE MODELOS
				UTILIZAR TÉCNICAS PREDICTIVAS	UTILIZAR TÉCNICAS DESCRIPTIVAS		
Recopilar e integrar las fuentes de datos existentes	Utilizar las técnicas de análisis exploratorio de datos	Detectar y tratar la presencia de valores atípicos (outliers)	Utilizar técnicas de reducción y aumento de la dimensión	Regresión y series temporales Análisis discriminante Métodos boyesianos Algoritmos genéticos Árboles de decisión Redes neuronales	Clustering y Segmentación Escalamiento Reglas de asociación y dependencia Análisis exploratorio Reducción de la dimensión	Intervalos de confianza Bootstrap Análisis ROC Evaluación de modelos	Visualización Simulación
Identificar y seleccionar las variables relevantes de los datos	Deducir la distribución de los datos, simetría y normalidad	Imputar la información faltante o valores perdidos	Aplicar técnicas de discretización y numerización				
Aplicar las técnicas de muestreo adecuadas	Analizar las correlaciones existentes en la información	Eliminar datos erroneos e irrelevantes	Realizar escalado simple y multidimensional				

Fuente: Pérez et al., 2007

2.2.5. CRISP-DM

Antes de CRISP-DM existían otras metodologías que eran ampliamente aceptadas por la comunidad científica dedicada a la ciencia de datos, en ese entonces denominada y más comúnmente conocida como minería de datos, sin embargo y de acuerdo a (Mardel, 2019) son diversas las metodologías que con el pasar del tiempo de han ido propuesto, entre ellas y más significativas: Sample, Explore, Modify, Model, Assess, SEMMA y CRISP-DM.

Habiendo escuchado tantas veces sobre CRISP-DM, entonces toca preguntarse qué es CRISP-DM o qué representa, por ejemplo (Plotnikova et al., 2022) define a CRISP-DM como un proceso estándar que captura un amplio rango de tareas recurrentes de minería de datos y entregables estructurados entorno a un ciclo del proyecto; también deja en claro que CRISP-DM es independiente de la industria, puesto que las organizaciones que desean usar esta metodología necesitan mayormente adaptar satisfacer sus requisitos específicos principales.

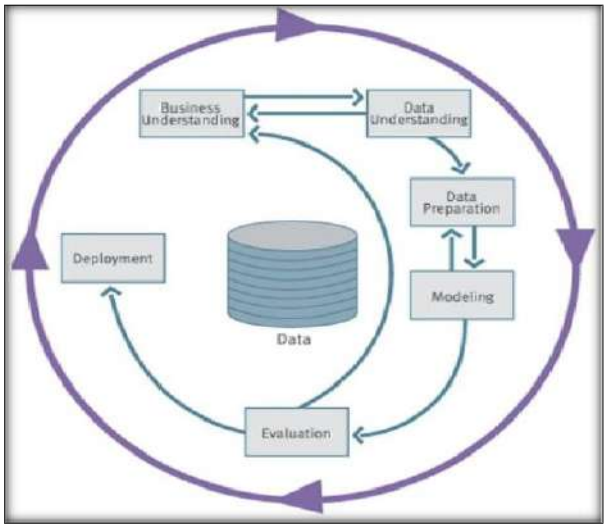
Según (Plotnikova et al., 2022) en términos de adopción y uso las propuestas basadas en KDD (SEMMA) han recibido una atención limitada en la academia y la industria por lo que la mayoría concuerda en elegir la metodología CRISP-DM debido a que SEMMA limita sus aplicaciones en otros entornos.

Podemos entender entonces por qué se prefiere mucho más a CRISP-DM y por qué es tan adaptado a los proyectos, en pocas palabras: guía a las personas para aplicar Data Mining en la práctica real de sistemas que están operativos. Claro que hay otro tipo de factores, en este sentido (Mardel, 2019) cita a las siguientes razones:

- CRISP-DM es de libre distribución.
- Constituye una de las guías de referencia más usadas a nivel mundial y sobre todo más confiables.
- Planifica tareas de manera jerárquica y dispone de métodos para reformar el proyecto de Data Mining en fases flexibles y cómodas para el usuario.
- Las tareas más generales pueden ser proyectadas a específicas y así se facilita la planificación y manipulación de trabajos de Data Mining.
- Entonces con lo último concluimos por qué CRISP-DM es la más aceptada por la amplia mayoría y las razones que la sitúan como la preferida por muchos y muchas.

Para llevar a cabo una correcta implementación de una minería de datos es necesario realizar pasos o tareas conforme se avance, para ello existen distintas metodologías enfocadas en estos tipos de proyectos. (Cazacu & Titan, 2020) resalta: “En cuanto a metodologías, CRISP-DM trae algo nuevo, una descripción general del ciclo de vida de un proyecto de minería de datos que contiene sus fases correspondientes, sus tareas y sus relaciones entre estas tareas”. Teniendo una estructura definida utilizada como guía, estas fases con bidireccionales lo que significa que algunas fases pueden revisar de forma total o parcial las fases previas o anteriores, nótese en la siguiente imagen”.

Imagen 29: Modelo metodología CRISP-DM



Fuente: Cazacu & Titan, 2020

2.3. MARCO CONCEPTUAL

2.3.1. PROCESO CRISP-DM

“La frase modelo de minería de datos (CRISP-DM) ha sido ampliamente aceptada durante las últimas dos décadas (…) CRISP-DM es un modelo genérico de proceso de minería de datos que proporciona una visión general de los ciclos de vida de los proyectos de minería de datos (…) Es considerado como el estándar de facto para el descubrimiento de conocimiento y proyectos de minería de datos” (Ayele, 2020).

Podemos entender entonces que, al ser el modelo más aceptado por la comunidad enfocada a ciencia de datos, debemos tener la idea de que ayuda enormemente a la exploración de orígenes de datos, valores de estos, resultados y productos obtenidos a través de la aplicación del proceso.

Como es bien sabido, el modelo CRISP-DM se estructura por fases, mismas que suman el total de seis; es importante recalcar que una fase tiene la posibilidad de volver a la fase anterior para su revisión o corrección.

FASE I: COMPRENSIÓN DEL NEGOCIO

“En esta etapa se lleva a cabo una comprensión de la sustancia de las actividades de minería de datos que se realizarán, así como la determinación de objetivos y la preparación de estrategias para lograr estos objetos”. (Giustin et al., 2022).

Se entiende entonces que, si esta fase no está bien entendida, se corre el riesgo de que los objetivos no sean comprendidos de tal forma que, si se omite algo tan fundamental como los objetivos, entonces tendremos problemas en un futuro para hacer la minería de datos o completar los obstáculos que deseamos resolver.

Lo explicado anteriormente no solo se queda ahí, sino que implica perder tiempo, ya que, al toparnos con los inconvenientes ya descritos, sabemos entonces que debemos retornar a fases anteriores e incluso al inicio, lo que en términos de economía empresarial nos puede jugar en contra o una muy mala pasada.

Hay que aclarar que estos objetivos que mencionamos tanto serán aquellos que van a resolver problemas en la empresa, pero usando la minería de datos, por tanto, los objetivos deben ser de minería de datos.

FASE II: COMPRENSIÓN DE LOS DATOS

“La comprensión de datos es el proceso de recopilar datos iniciales y estudiarlos para comprender lo que pueden hacer”. (Giustin et al., 2022).

Deduje entonces que, si la recolección de datos inicia en este punto, demandará desde ya un esfuerzo significativo para las distintas operaciones que se tengan que realizar, por ejemplo: se tiene que describir a los datos y crear conjuntos o volúmenes de ellos para identificarlos y saber el formato de estos.

En esta fase también, se incluye el tema de exploración de datos. De cierta forma, y como es bien sabido, hay que identificar de alguna manera el patrón de los datos en términos estadísticos.

FASE III: PREPARACIÓN DE LOS DATOS

“La preparación de datos consiste en crear una nueva base de datos que se utilizará para el proceso de minería de datos. Los datos obtenidos todavía están en forma de datos no estructurados, en los que el contenido de los datos todavía contiene ruido. Por lo tanto, en el proceso de preparación de datos, se lleva a cabo la limpieza de datos, incluida la eliminación de datos duplicados y la corrección de errores en los datos”. (Giustin et al., 2022).

Para preparar los datos también debemos modelar, elegir una técnica para ello y posteriormente hacer la limpieza de los datos, generar alguna variable adicional e integrar diversos orígenes de datos con un determinado cambio de formato.

Antes de todo lo descrito anteriormente, hay que apoyarnos en la selección de datos, es decir, si elegimos una cantidad determinada de datos debemos corregirlos, limitarlos en volumen u observar

FASE IV: MODELADO

“El modelado es la etapa de selección y aplicación de varias técnicas de modelado y algunos de los parámetros se ajustarán para obtener el valor óptimo”. (Giustin et al., 2022).

Como se describió en la fase anterior, en esta también, pero siendo una actividad mucho más centrada en lo que respecta a modelado, debemos elegir una técnica de modelado que sea la apropiada para resolver el objetivo principal del proyecto, ya que, existen demasiadas en el mercado y no todas satisfacen las necesidades de los objetivos finales.

Finalmente, se tiene que construir el modelo, en este punto hay un aspecto importante y es que, si a la selección de la técnica implicará elegirla con todo y sus parámetros, lo que al final influye en las

características del modelo que se va a generar. Por último, evaluar el modelo.

FASE V: EVALUACIÓN

"La evaluación del modelo es la etapa de determinar si el modelo construido está de acuerdo con los objetivos planteados en la fase inicial. En esta etapa, la evaluación del modelo se realiza usando una matriz de confusión. La matriz de confusión es un método que generalmente se usa para realizar cálculos de precisión sobre conceptos de minería de datos o sistema de soporte de decisiones". (Giustin et al., 2022).

Se entiende entonces que se deben evaluar los resultados, pero en relación con los objetivos del negocio y encontrar que razones por las que el modelo generado anteriormente es deficiente o no.

Como último punto, si se tienen todas las evaluaciones correctas y estas han arrojado resultados que indican que han sido satisfactorios para los objetivos, entonces toca determinar las próximas etapas a seguir.

FASE VI: DESPLIEGUE

"Elabora un informe sobre el conocimiento obtenido o el reconocimiento de patrones en el proceso de minería de datos que se presenta en forma de gráficos o descripciones que son fáciles de entender". (Giustin et al., 2022).

Como se lee líneas arriba, un proyecto de minería de datos no termina solamente en la implantación del modelo, sino que hay que, primero, planear como se implantará y luego gestionar todo aquel proceso que sirva para monitorear y mantener los modelos obtenidos.

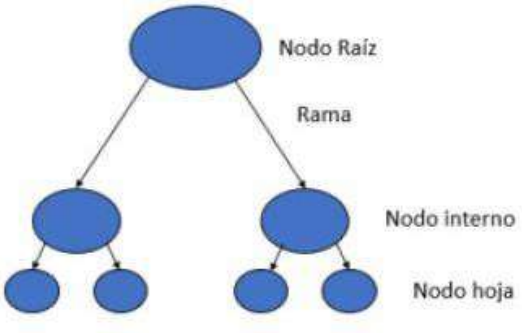
2.3.2. ÁRBOLES DE DECISIÓN

Una de las técnicas de minería de datos que más se usan es el árbol de decisión y de acuerdo con lo que dice (Romero, 2020) estos árboles de decisiones son un método que usa la regresión y clasificación, lo que quiere decir que segmentar el espacio de predictores en un número de regiones simples.

Para poder hacer una predicción usando árboles de decisiones, en principio, debemos tener una observación y después usar el promedio en caso de regresión o tomar la mayor clase del conjunto de datos que hemos usado para el entrenamiento en la región si es para tareas de clasificación.

Al escuchar el nombre árbol, automáticamente cae en cuestión por qué se le llama así y es que es una estructura similar a la de un árbol, presenta un nodo interno que representan a los atributos, ramas que representan a reglas de decisión y cada nodo hoja viene a representar un resultado; esto se representa mejor en la siguiente imagen.

Imagen 30: Partes de un árbol de decisión



Fuente: Romero, 2020

Autores como (Urbina-Nájera & Méndez-Ortega, 2022), definen los árboles de decisión de manera similar, en su caso los conceptualiza como una técnica de aprendizaje supervisados, fácil de implementar y muy útil, compuesta por un único nodo inicial y debajo de otros árboles independientes que indican la predicción de atributos. Cabe destacar que los árboles de decisión están ubicados dentro de una rama de Machine Learning llamado aprendizaje simbólico, en el que también podemos encontrar modelos de reglas de decisión fuertemente relacionadas con los árboles.

2.3.3. BINNING

Antes de entrar en profundidad con lo que es el Binning, también conocido como Data Binning, tenemos que adentrarnos al campo qué llama o requiere el uso de esta técnica; este campo es denominado el de los “datos ruidosos”; el ruido en los datos, tal como explica (Muñoz, 2022): “Es un procedimiento que agrupa grupos de valores que forman parte de un conjunto de datos, teniendo como principal beneficio la reducción de costes computacionales”.

Es entonces que ingresa el Binning, esta técnica es una de las técnicas de suavizado o agrupamiento, es decir que el valor de los datos será ordenado consultando a otros valores que lo rodean, también llamados vecindad o vecinos. Para (Minitab, 2019) “Se denomina al Data Binning como una categorización o técnica de ella, es una manera de simplificar y comprimir una columna de datos”; por tanto y para nuestro proyecto en particular se va a poder reducir el número de niveles; del mismo modo se simplificará y

comprimirá aquella columna de datos reduciéndola al número de valores o niveles posibles representados en los datos.

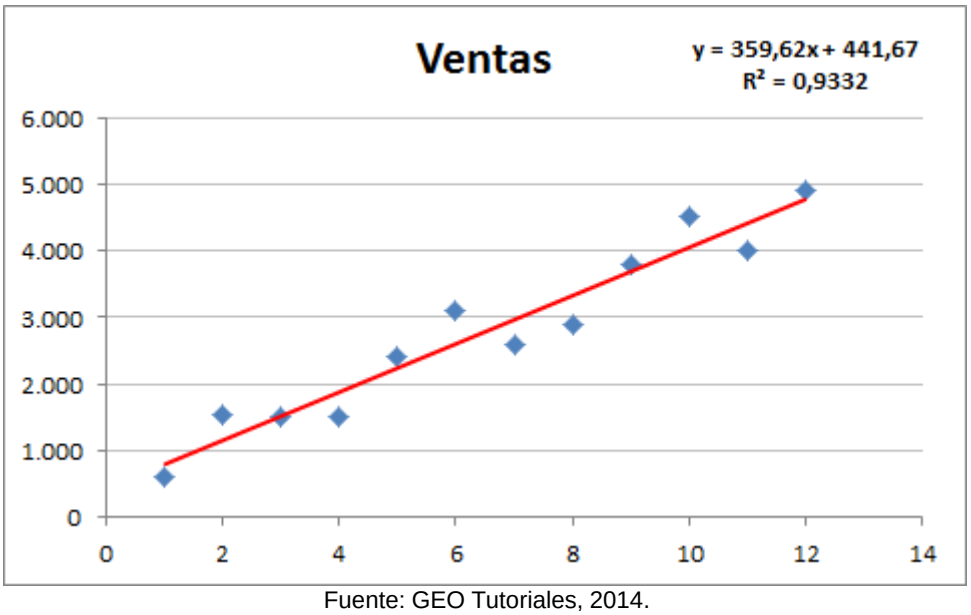
Podemos indicar que para el usar Data Binning existen muchas ventajas, entre ellas está:

- Prepara los datos a binarios, ya que, las máquinas de aprendizaje (p. ej., SVM) requieren de ellos.
- Ofrece protección contra errores menores en los datos.
- Ofrece protección contra valores atípicos.
- Puede aplicarse a datos continuos o categóricos y para también para datos numéricos o textuales.

Alguna de las técnicas del Binning suelen ser las siguientes:

- **Regresión**: En palabras de (Muñoz, 2022): “El Binning de datos puede ser realizado mediante el modelo general de regresión, el cual consiste en analizar un dataset y estimar las relaciones existentes entre variables. Son estas relaciones entre variables a lo que se le denomina regresión.” Es claro la postura del autor con respecto a ello y es que la regresión lineal consiste en encontrar la mejor línea para ajustar dos variables de forma que un atributo pueda usarse para predecir otro. Por otro lado, tenemos a la regresión lineal múltiple que es una extensión de la lineal, sin embargo, intervienen más de dos atributos y los datos se ajustarán a una superficie multidimensional.

Imagen 31: Regresión lineal en un ejemplo de ventas

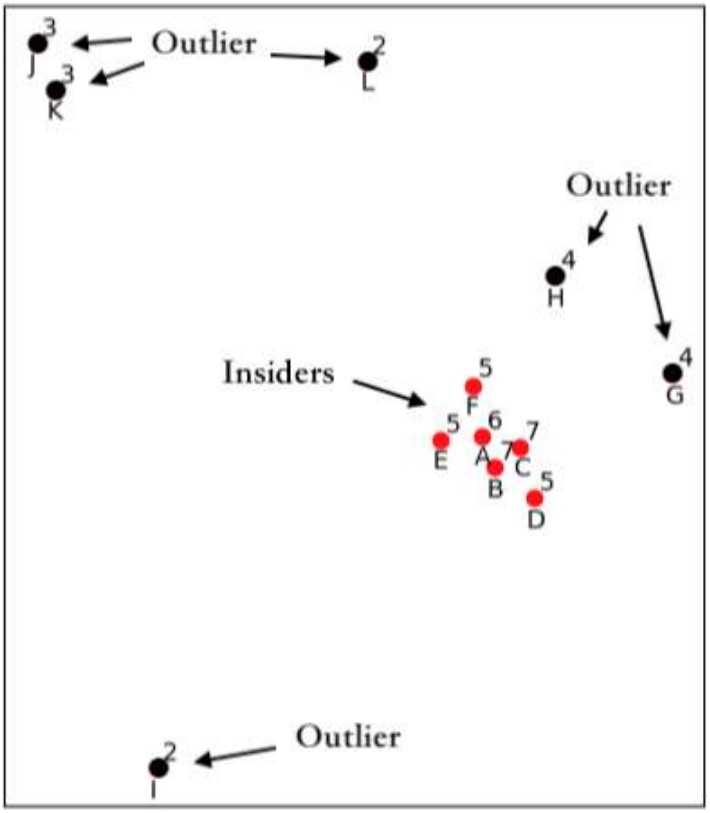


Fuente: GEO Tutoriales, 2014.

- **Análisis de valores atípicos:** De acuerdo con (Leslie & Salazar, 2022): “Durante el preprocesamiento de datos, se reconocen valores que escapan del rango normal de donde se concentran la mayoría de características”. Tal cual el autor expresa, de forma intuitiva aquellos

valores que quedan fuera de estos grupos mencionados serán los valores atípicos; hay una amplia variedad de técnicas para la eliminación de estos valores y una de ellas es el Binning.

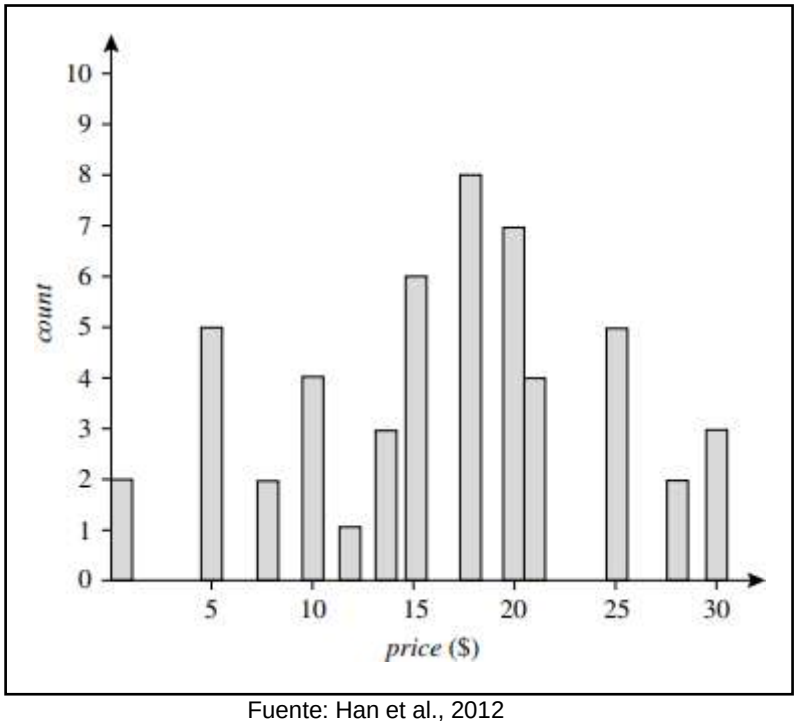
Imagen 32: Ejemplo de agrupamiento con outliers



Fuente: Zhang, W. & Zhang, G., 2019

Por último, en cuanto a lo que respecta al proyecto en sí, se usarán histograma que, para (Knuth, 2019) contextualizan a los histogramas de la siguiente manera: “Los histogramas son estimadores de densidad no paramétricos, que continúan usándose de manera ubicua. Los cuales utilizan Binning para estimar cantidades de resumen a partir de modelos de densidad probabilísticos los cuales dependen de la elección del número de contenedores.”. Por tanto, es requerido saber de qué manera funciona un histograma que usa Data Binning, en este caso, para un atributo X, el histograma dividirá la distribución de datos de X en subconjuntos separados, denominados cubos o contenedores; si cada cubo representa un solo par atributo-valor-frecuencia, entonces los cubos se denominarán cubos únicos. En la mayoría de los casos, los cubos representan rangos continuos para el atributo en cuestión.

Imagen 33: Histograma de precios que usa cubos individuales



Fuente: Han et al., 2012

2.3.4. REGLAS DE ASOCIACIÓN

Para llevar a cabo la correcta implementación de minería de datos, es necesario cumplir con todas las fases para su procesamiento, y así obtener buenos resultados. Las reglas de asociación forman parte de la última fase, siendo esta una técnica muy importante enfocada en descubrir características comunes o similares del conjunto de datos para poder, como su mismo nombre lo dice, asociarlas.

(Pérez, 2018) define las reglas de asociación como la técnica que analiza la base de datos en función a las características que ocurren con frecuencia juntas, descubriendo relaciones de asociación o correlación el cual se denomina un patrón.

Se plantea un problema para un mejor entendimiento, siendo este: Identificar el conjunto de ítems que son adquiridos en conjunto. Se aplica reglas a la forma {fideos, queso} -> {salsa]. En conclusión, si se compra fideos y queso, es probable que también se compre salsa.

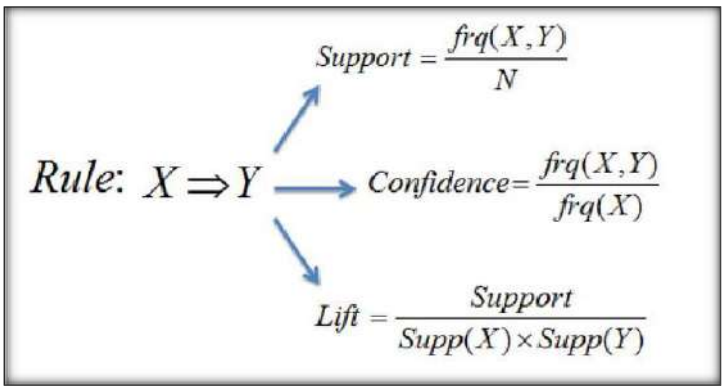
Se puede aplicar la teoría en la siguiente forma general: X -> Y (X y Y son conjunto de ítems). X es denominado antecedente y Y consecuente.

(Monteserin, 2018) define métricas de soporte y confianza en las reglas de asociación, y las interpreta. Una regla de bajo con bajo soporte puede haber aparecido por casualidad y una regla con baja confianza es probable que no

exista relación entre X y Y. El objetivo de las reglas de asociación viene a ser encontrar valores con alto nivel de soporte y confianza para un buen análisis.

(Pérez, 2018) señala tres indicadores: lift, soporte y confianza. El lift nos indica si realmente existe o no la asociación y si es positiva o negativa. Además, indica que, al no tener buenos resultados, estos se clasifican en tres tipos: Factibles, triviales e imposibles.

Imagen 34: Fórmulas de soporte, confianza y tasa de soporte observada



Fuente: JayJay, 2018

La minería de datos y las reglas de asociación hacen uso de algoritmos predictivos para descubrir patrones de comportamiento ocultos. Existen distintos algoritmos, pero todos generan el mismo conocimiento, lo que los hace diferentes según (Monteserin, 2018): “Vienen a ser la forma de carga de los datos, el tiempo de procesamiento, tipos de atributos (numéricos, categóricos), forma en la que los itemsets son generados y la estructura de datos utilizada”.

ALGORITMO A PRIORI

(Sáenz et al., 2017) define el proceso del algoritmo Apriori como la obtención de conjunto de ítems frecuentes los cuales los agrupa en tamaño 1, tamaño 2 y así sucesivamente, hasta que no se encuentren conjuntos de ítems con soporte mayor al soporte mínimo. En el siguiente ejemplo se supone un conjunto de ítems {a}, {b}, {c}, {d}, {e} y se buscarán acorde a la siguiente tabla.

Tabla 14: Ejemplo conjunto de ítems

Conjunto de ítems	Núm. De transacciones
Conjunto de un solo item	{a}, {b}, {c}, {d}, {e}
Conjunto de dos items	{a,b}, {a,c}, {a,d}, {a,e} {b,c}, {b,d}, {b,e} {c,d}, {c,e}

	{d,e}
Conjunto de tres items	{a,b,c}, {a,b,d}, {a,b,e}, {a,c,d}, {a,c,e}, {a,d,e} {b,c,d}, {b,c,e}, {b,d,e} {c,d,e}
Conjunto de cuatro items	{a,b,c,d}, {a,b,c,e} {a,b,d,e} {a,c,d,e} {b,c,d,e}
Conjunto con cinco items	{a,b,c,d,e}

Fuente: Sáenz et al., 2017

El algoritmo Apriori viene a ser fácil de implementar, entender y pueden aplicarse a conjuntos de elementos grandes. Sin embargo, es necesario encontrar un gran número de reglas candidatas, desde el punto de vista informático el cálculo del soporte también viene a ser complicado porque tiene que recorrer toda la base de datos. Para llevar a cabo el algoritmo Apriori se consideran los siguientes pasos con el fin de acelerar el proceso en caso el conjunto de datos sea más grande (Gonzalez, 2021).

- Paso 1. Establecer el valor mínimo para el soporte y la confianza.
- Paso 2. Extraer los subconjuntos con valor de soporte superior a un umbral mínimo.
- Paso 3. Seleccionar todas las reglas de los subconjuntos con un valor de confianza superior al umbral mínimo
- Paso 4. Ordenar las reglas por orden descendente de lift.

ALGORITMO PARTITION

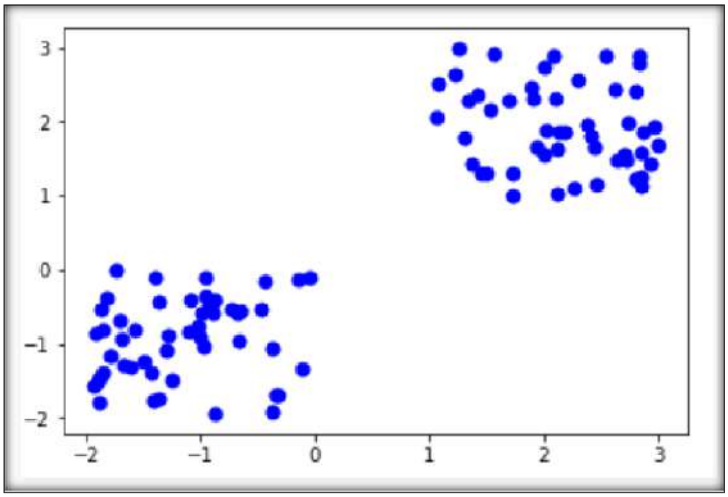
Conocido como algoritmo K-Mean, este método de particionamiento agrupa la información en clústeres. (Chandra, 2020) señala que este algoritmo clasifica la información en múltiples grupos basado en similitud de características de la data. Los analistas son los que especifican el número de clústeres que se generarán para el agrupamiento.

Para llevar a cabo el algoritmo de clasificación K-Mean se minimizan la suma de distancia entre objetos y el centro del clúster. Se aplica el siguiente método:

1. Asignar aleatoriamente K objetos del conjunto de datos (D) como centros de clústeres (C)
2. Asignar objetos más similares entre sí en función a los valores medios.
3. Actualizar clúster, recalculando la media de cada clúster con los nuevos valores asignados.

4. Repetir.

Imagen 35: Agrupación de K-mean



Fuente: Chandra, 2020

ALGORITMO ECLAT

El algoritmo Eclat, según (Báez et al., 2018) “Tiene como función agrupar los ítems más frecuentes y luego emplear algoritmos más eficientes para agrupar los ítems más frecuentes contenidos en cada grupo. Para este agrupamiento se proponen dos métodos después de descubrir los conjuntos frecuentes, el primero, por clases de equivalencia (agrupa los itemsets que tienen el primer ítem igual) y segundo, por la búsqueda de cliques maximales (agrupa los ítems por aquellos que forman cliques maximales).”

Es una versión más eficiente y escalable del algoritmo Apriori, con la única diferencia que Eclat funciona de manera vertical como la búsqueda en profundidad primero de un gráfico, mientras que el algoritmo Apriori funciona en sentido horizontal imitando la búsqueda primero en amplitud de un gráfico. Este enfoque vertical del algoritmo Eclat lo hace mucho más rápido.

Para cada ítem almacena en una lista en que transición aparece cada ítem. Ejemplo de algoritmo Eclat (Gupta, 2021).

Imagen 36: Ejemplo matriz booleana

Transaction Id	Bread	Butter	Milk	Coke	Jam
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

Fuente: Gupta, 2021

Imagen 37: K=1, soporte mínimo=2

Articulo	Tidset
Pan de molde	{T1, T4, T5, T7, T8, T9}
Mantequilla	{T1, T2, T3, T4, T6, T8, T9}
Leche	{T3, T5, T6, T7, T8, T9}
Coca	{T2, T4}
Mermelada	{T1, T8}

Fuente: Gupta, 2021

A partir de aquí llamamos de forma recursiva a la función hasta que no se pueda combinar más pares de artículo-tidset.

Imagen 38: K=2

Articulo	Tidset
{Pan con mantequilla}	{T1, T4, T8, T9}
{Pan, Leche}	{T5, T7, T8, T9}
{Pan, Coca Cola}	{T4}
{Pan, Mermelada}	{T1, T8}
{Suero de la leche}	{T3, T6, T8, T9}
{Mantequilla, Coca Cola}	{T2, T4}
{Mantequilla, Mermelada}	{T1, T8}
{Leche, mermelada}	{T8}

Fuente: Gupta, 2021

Imagen 39: K=3

Articulo	Tidset
{Pan, mantequilla, leche}	{T8, T9}
{Pan, mantequilla, mermelada}	{T1, T8}

Fuente: Gupta, 2021

Imagen 40: K=4

Articulo	Tidset
{Pan, mantequilla, leche, mermelada}	{T8}

Fuente: Gupta, 2021

Nos detenemos en k = 4 porque no hay más pares artículo-tidset para combinar.

Imagen 41: Conclusión reglas de conjunto de datos

Artículos comprados	Productos Recomendados
Pan de molde	Mantequilla
Pan de molde	Leche
Pan de molde	Mermelada
Mantequilla	Leche
Mantequilla	Coca
Mantequilla	Mermelada
Pan y mantequilla	Leche
Pan y mantequilla	Mermelada

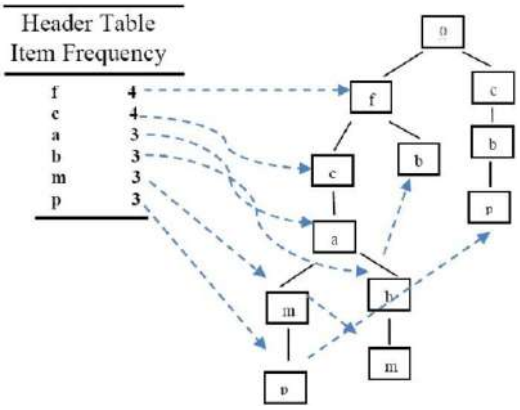
Fuente: Gupta, 2021

ALGORITMO FP GROWTH

(Aekwarangkoon & Thanathamathee, 2022) argumenta: "(…) Existen varios algoritmos para encontrar reglas de asociación, el más conocido y usado es el algoritmo FP-Growth. El algoritmo FP-Growth usa el principio de dividir y conquistar el árbol FP para establecer reglas de asociación comunes. Esta técnica puede establecer rápidamente una regla de relación, la limitación es que los algoritmos de uso de datos deben esar en forma binaria y requiere mucha memoria cuando el árbol es grande (…)".

El algoritmo FP Growth transforma la base de datos en una estructura llamada FP-Trees en dónde toma distintas transacciones repetidamente, compartiendo nodos para aumentar su soporte, para la construcción del árbol con ítems en forma descendente. Se toma en cuenta los nodos más profundos, junto con el soporte el cual determina la ubicación en el árbol, y se construye los patrones más frecuentes.

Imagen 42: Ejemplo de implementación de algoritmo FP-Growth

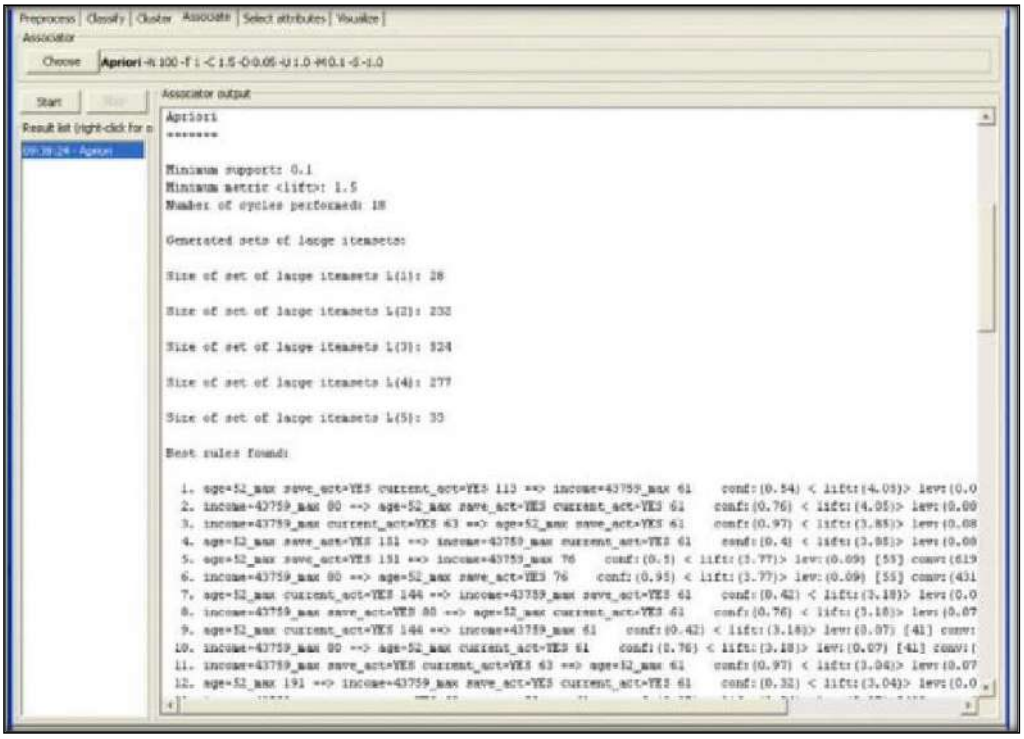


Fuente: Mansour et al., 2012

REGLAS DE ASOCIACIÓN en WEKA

Tomando en cuenta que la interfaz Explorer de WEKA se ha mostrado anteriormente. (Soler, 2017) muestra la implementación de reglas de asociación en WEKA de la siguiente manera. Partiendo de una interfaz ya ejecutada utilizando el algoritmo a priori, las reglas de asociación se mostrarán de la siguiente manera.

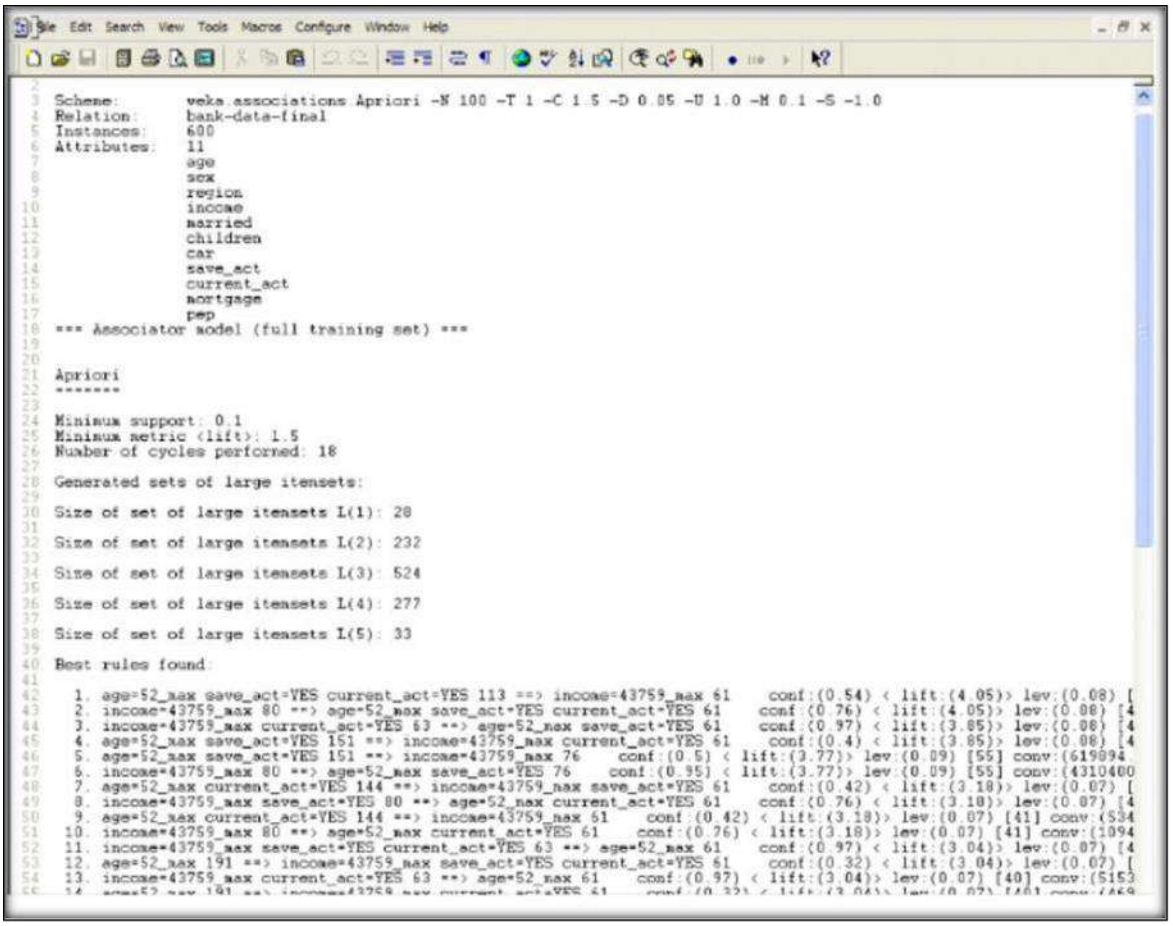
Imagen 43: Resultados del algoritmo A priori



Fuente: Soler, 2017

En el panel de la izquierda de lista de resultados se puede llevar a cabo la aplicación del algoritmo n veces, y las ejecuciones aparecerán en dicho panel, con parámetros diferentes, al hacer clic se podrán observar los resultados de cada lista en el panel de la derecha.

Imagen 44: Resultados de ejecución del algoritmo A priori (WEKA)



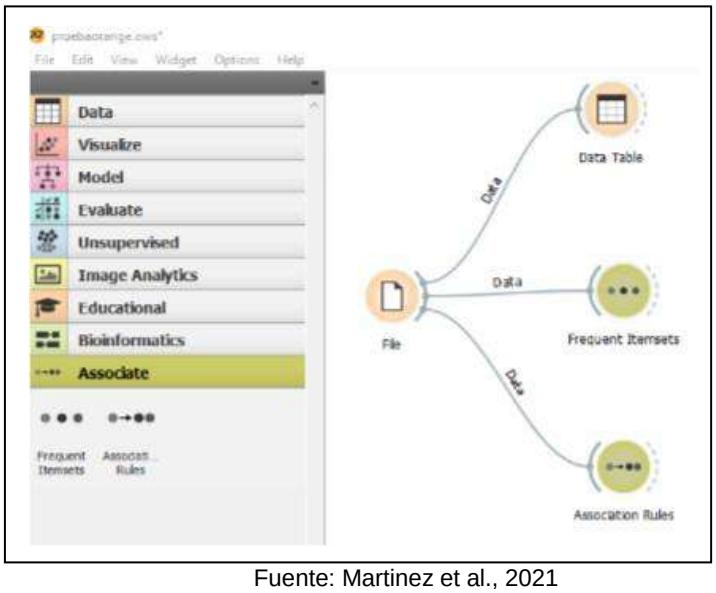
Fuente: Soler, 2017

Los resultados se interpretarán en base a los datos cargados tomando en cuenta sus características principales tales como confianza, soporte y lift.

REGLAS DE ASOCIACIÓN en ORANGE

Según el artículo de (Martinez et al., 2021) muestra que el software Orange tiene una interfaz amigable y práctica a nivel gráfico acompañado de los widgets, de esta forma se muestra el siguiente flujo de trabajo.

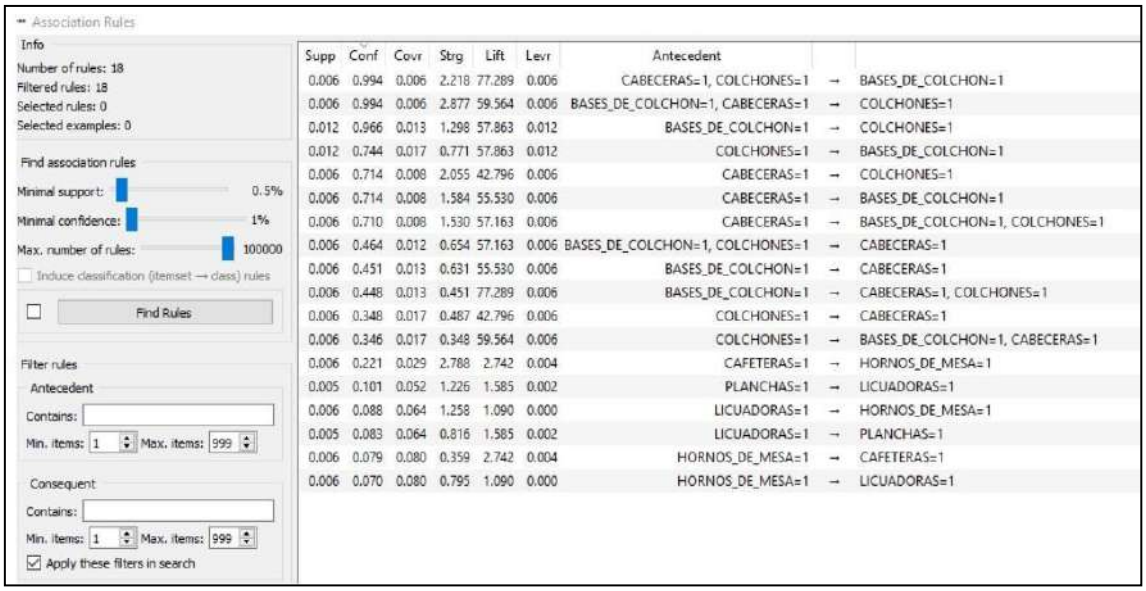
Imagen 45: Estructura de nodos para reglas de asociación (ORANGE)



Fuente: Martinez et al., 2021

El proceso a llevar a cabo parte desde leer el archivo de nuestros datos y conectarlas hacia las reglas de asociación. Con motivo de mostrar un ejemplo de las reglas de asociación, en la siguiente imagen se puede visualizar 18 reglas obtenidas producto de la selección de un soporte mínimo de 0.05% y una confianza mínima de 1%, usando FP-Growth.

Imagen 46: Resultados de reglas de asociación (ORANGE)



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.006	0.994	0.006	2.218	77.289	0.006	CABECERAS=1, COLCHONES=1	→	BASES_DE_COLCHON=1
0.006	0.994	0.006	2.877	59.564	0.006	BASES_DE_COLCHON=1, CABECERAS=1	→	COLCHONES=1
0.012	0.966	0.013	1.298	57.863	0.012	BASES_DE_COLCHON=1	→	COLCHONES=1
0.012	0.744	0.017	0.771	57.863	0.012	COLCHONES=1	→	BASES_DE_COLCHON=1
0.006	0.714	0.008	2.055	42.796	0.006	CABECERAS=1	→	COLCHONES=1
0.006	0.714	0.008	1.584	55.530	0.006	CABECERAS=1	→	BASES_DE_COLCHON=1
0.006	0.710	0.008	1.530	57.163	0.006	CABECERAS=1	→	BASES_DE_COLCHON=1, COLCHONES=1
0.006	0.464	0.012	0.654	57.163	0.006	BASES_DE_COLCHON=1, COLCHONES=1	→	CABECERAS=1
0.006	0.451	0.013	0.631	55.530	0.006	BASES_DE_COLCHON=1	→	CABECERAS=1
0.006	0.448	0.013	0.451	77.289	0.006	BASES_DE_COLCHON=1	→	CABECERAS=1, COLCHONES=1
0.006	0.348	0.017	0.487	42.796	0.006	COLCHONES=1	→	CABECERAS=1
0.006	0.346	0.017	0.348	59.564	0.006	COLCHONES=1	→	BASES_DE_COLCHON=1, CABECERAS=1
0.006	0.221	0.029	2.788	2.742	0.004	CAFETERAS=1	→	HORNOS_DE_MESA=1
0.005	0.101	0.052	1.226	1.585	0.002	PLANCHAS=1	→	LICUADORAS=1
0.006	0.088	0.064	1.258	1.090	0.000	LICUADORAS=1	→	HORNOS_DE_MESA=1
0.005	0.083	0.064	0.816	1.585	0.002	LICUADORAS=1	→	PLANCHAS=1
0.006	0.079	0.080	0.359	2.742	0.004	HORNOS_DE_MESA=1	→	CAFETERAS=1
0.006	0.070	0.080	0.795	1.090	0.000	HORNOS_DE_MESA=1	→	LICUADORAS=1

Fuente: Martinez et al., 2021

En ella se puede configurar las reglas de asociación (mínimo de soporte, mínimo de confianza y máxima cantidad de reglas a mostrar), el resultado se aprecia en la sección de la derecha a espera de su interpretación de acuerdo con el dataset utilizado.

MATEMÁTICA Y ESTADÍSTICA DETRÁS DE REGLAS DE ASOCIACIÓN

Los parámetros de las reglas de asociación según (Sáenz et al., 2017) vienen a ser el soporte y confianza. Estos parámetros nos ayudarán a conocer y seleccionar las reglas de asociación generadas por el algoritmo, dependiendo de su importancia de estas en base a sus valores.

i. Soporte

Frecuencia en la cual el ítem se encuentra en las transacciones dividido entre el número de transacciones.

$$Soporte\ (A) = \frac{Nro\ de\ transacciones\ que\ contienen\ en\ el\ ítem\ A}{Nro\ de\ transacciones\ de\ la\ base\ de\ datos}$$

Para obtener el soporte de una regla de asociación, con ejemplo de A → B, se obtiene:

$$Soporte\ (A \longrightarrow B) = \frac{Nro\ de\ transacciones\ que\ contienen\ en\ el\ ítem\ A\ y\ B}{Nro\ de\ transacciones\ de\ la\ base\ de\ datos}$$

ii. Confianza

La medida confianza de una regla de asociación (A → B) es la división entre el soporte de la regla de decisión ante el soporte del antecedente de la regla de decisión.

$$Conf\ (A \longrightarrow B) = \frac{Soporte\ (A, B)}{Soporte\ (B)}$$

iii. Lift

Dada una regla L → R, el lift es la relación entre la probabilidad de que L y R se presenten juntas, además, la multiplicación de las dos probabilidades individuales. La siguiente ecuación.

$$Lift(L \longrightarrow R) = \frac{\Pr(L, R)}{\Pr\ (L) x\ \Pr(R)}$$

Si lift = 1, entonces L y R son independientes. Mientras mayor sea este valor la probabilidad de que L y R existan juntos no es una ocurrencia aleatoria, si no que existe una relación entre ellas.

2.3.5. REDES BAYESINAS

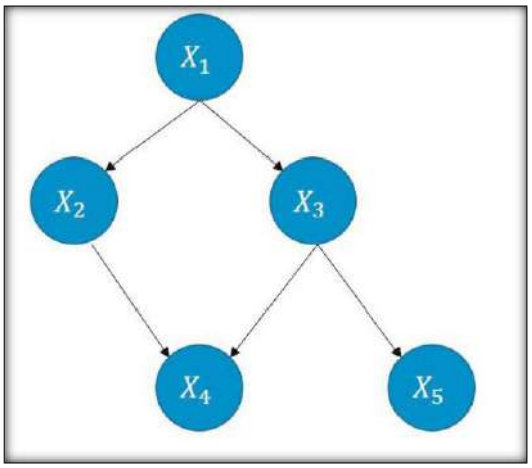
En palabras de (Wang, 2022): “Las redes bayesianas son el modelo gráfico para representar relaciones de probabilidad casual entre variables multiatributo (…) se trata de una estructura de red basada en grafos acíclicos dirigidos para representar las dependencias entre atributos y usa tablas de probabilidad condicional para describir las distribuciones de probabilidad conjuntos (…) ”. De tal modo podemos observar a través de la representación una codificación formal de la distribución de probabilidad conjunta para su domino, sin embargo, incluye una estructura cualitativa orientada al humano para que de esta manera sea mucho más sencillo la comunicación entre usuario y sistemas que incorporará el modelo probabilístico.

(Soler-Flores et al., 2019) definen y explican que una red bayesiana consta de lo siguiente:

- Una red bayesiana es un gráfico acíclico dirigido.
- Cada nodo representa una variable aleatoria.
- Las dependencias entre variables están codificadas en la estructura del gráfico de acuerdo con el criterio de separación.
- Para una red con N variables, la ecuación es la siguiente:

$$p(x1, x2, \dots, Xn) = \prod_{i=1}^{n} p(Xi \mid pa(Xi))$$

Imagen 47: Grafo dirigido con un ciclo de retroalimentación



Fuente: Soler-Flores et al., 2019

(Soler-Flores et al., 2019) indica que: “Las redes bayesianas aprovechan la expresividad del gráfico para automatizar el proceso de desarrollo probabilístico. El resultado combina la teoría de grafos y la probabilidad, esta unión permite el modelado eficiente del aprendizaje automático mediante parámetros modelados por distribuciones beta, distribución de Dirichlet y la inferencia de la evidencia disponible”. El estudio de esta técnica brinda una visión general del problema al aprendizaje estadístico y a la minería de datos.

Tabla 15: Estimación de parámetros

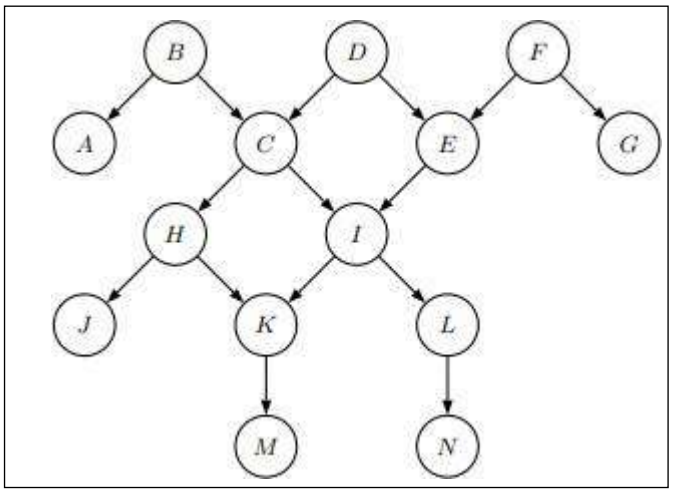
Estimador	Expresión
Máxima verosimilitud. Multinomial	$\theta_k^* = \frac{N_k}{N}$
Estimación bayesiana. Dirichlet	$\theta_k^* = \frac{N_k + a_k}{N + \sum_{i=1}^{r} a_i}$

Fuente: Soler-Flores et al., 2019

MANTO DE MARKOV

En palabras de (Del Río Cárdenas, 2019): “(…) el manto de Markov de un nodo objetivo consiste de sus nodos padres, hijos directos y padres de los hijos (esposos); siendo este nodo independiente del resto de nodos de la red dado su manto de Markov.”

Imagen 48: Ejemplo de grafo dirigido



Fuente: Jensen & Nielsen, 2007

Para entender de mejor manera lo explicado anteriormente, la manta de Markov para I vendría a hacer {C, E, H, K, L}.

SEPARACIÓN-D

Haciendo referencia a (Mateos, 2021) quien dice que: “Se le llama separación-d al criterio que se sigue cuando existe una desconexión de dos nodos debido a un tercero”. Por ejemplo, si A, B y C son 3 nodos, o subconjuntos, por tanto, C provoca la separación-d de A con respecto a B que se podría definir matemáticamente de la siguiente forma:

$$\mathrm{A\ M\ G\ B|C}$$

Las redes bayesianas tienen una parte muy importante dentro del modelo que se denomina inferencia; para (Ramírez, 2020): “(…) la inferencia es el proceso de cálculo de la probabilidad una vez conocidos los valores que toman otras variables de la red, es decir, cuando se introduce una determina evidencia (…)”.

De la cita anterior podemos entender con el siguiente ejemplo: sea el conjunto o el total de variables en la red y el conjunto X de variables observadas, de las cuales de hace un proceso de ingestión de evidencias (X = x), donde Y es el conjunto de variables que no se observan, ahí se encontrará un subconjunto S donde se encontrarán las variables a ingerir y T aquellas variables de interés.

(Ramírez, 2020) define el proceso de inferencia como el cálculo de:

$$P(S|X = x) = \frac{P(S, X = x)}{P(X)}$$

2.3.6. SOFTWARE WEKA

La plataforma de WEKA viene a ser un software libre enfocado en el aprendizaje automático y MD. En su versión 3.8.5, la cual se utiliza en el proyecto, nos permite trabajar y elegir entre cinco opciones.

Imagen 49: Versión 3.8.5 de software Weka

Fuente: Elaboración propia

En la siguiente tabla se explicarán brevemente cuatro de ellas según (Molina & García, 2006):

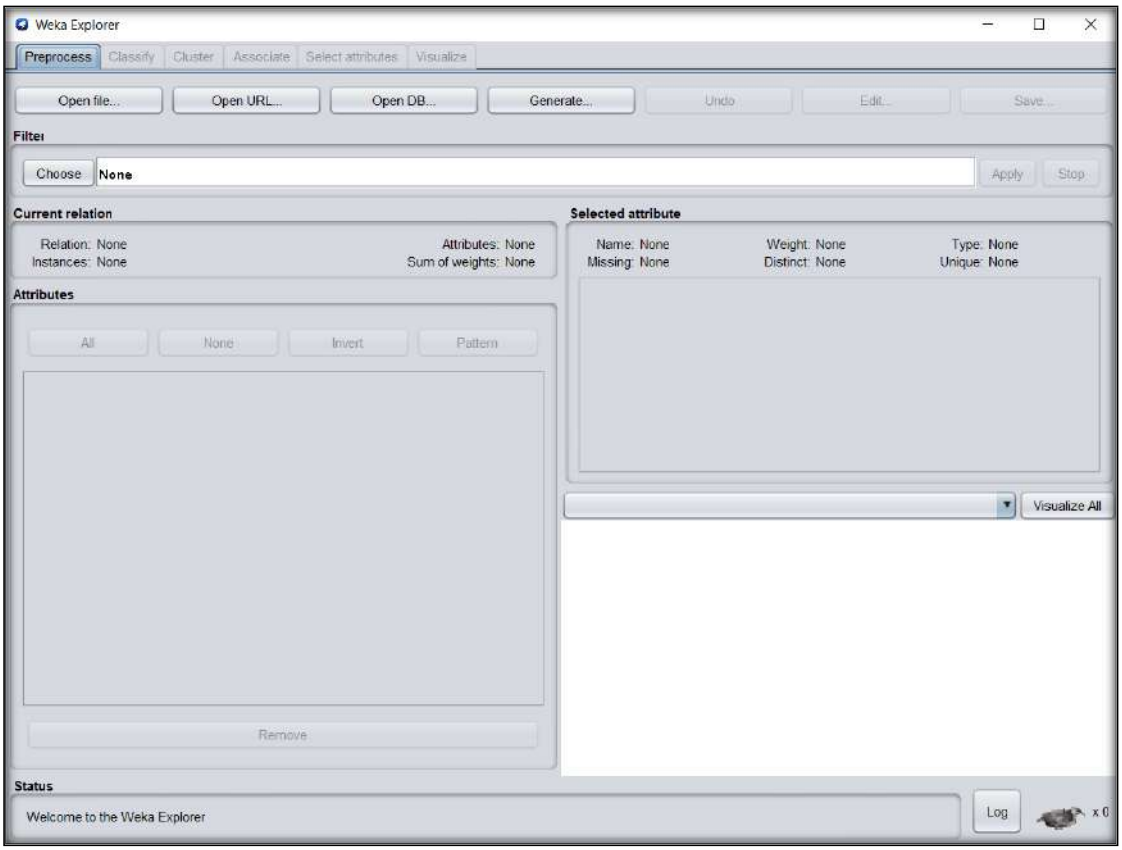
Tabla 16: Opciones de software Weka

Explorer	En esta opción se lleva a cabo la ejecución de los distintos algoritmos de análisis que WEKA ofrece, el proceso se da con algún fichero como entrada, y será la opción que se utilizará para el proyecto.
Experimenter	Define experimentos más complejos, compara estadísticamente los resultados de uno o varios algoritmos ejecutados en uno o varios ficheros de entrada.
KnowledgeFlow	Lleva a cabo las mismas acciones de la opción Explorer con diferencia a que ofrece una configuración gráfica.
Simple CLI	CLI significa “Command-Line Interfaz” y es utilizada como ventana de comandos java para ejecutar clases de weka.

Fuente: Molina et al., 2006

La opción Explorer se muestra de la siguiente manera en WEKA:

Imagen 50: Interfaz Weka Explorer



Fuente: Elaboración propia

Pondremos todo el enfoque en la opción Explorer de WEKA para llevar a cabo el proyecto, la anterior imagen muestra 6 pestañas en la parte superior que corresponden a los distintos tipos de operaciones que pueden llevarse a cabo sobre los datos. Explicados a continuación en la siguiente tabla:

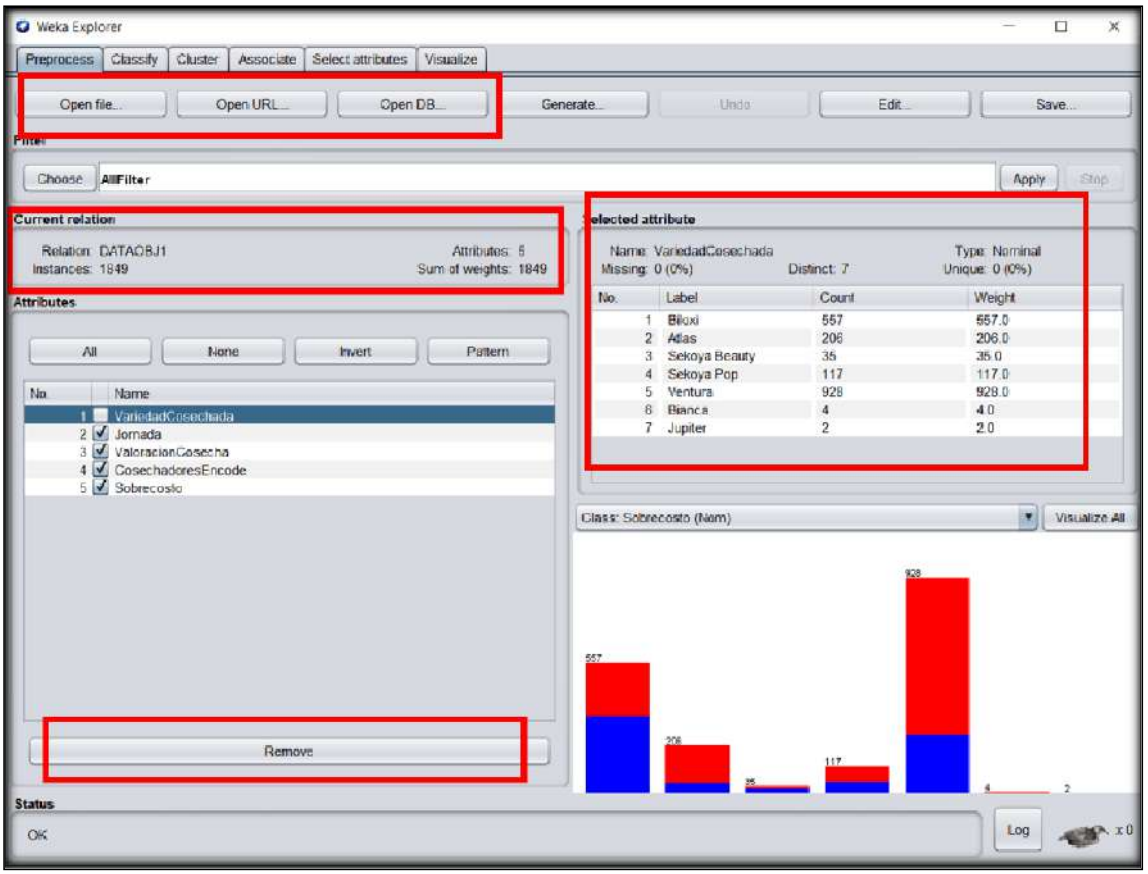
Tabla 17: Operaciones en la opción Weka Explorer

PREPROCESS	Selección de fuente de datos y filtrado
CLASIFY	Entrenar el modelo, determinar precisión
CLUSTER	Algoritmos de agrupación
ASSOCIATE	Algoritmos de búsqueda de asociación
SELECT ATTRIBUTES	Búsqueda supervisada de subconjuntos de atributos representativos
VISUALIZE	Presentación gráfica de 2D

Fuente: Elaboración propia

A continuación, se explicará detalladamente cada una de ellas, comenzando por el preprocesado de los datos, “es considerada la primera parte antes de realizar alguna otra operación, ya que es en esta parte en donde se identifican los datos para llevar a cabo cualquier análisis”(Molina & García, 2006).

Imagen 51: Interfaz y funciones del Pre-process de Weka Explorer

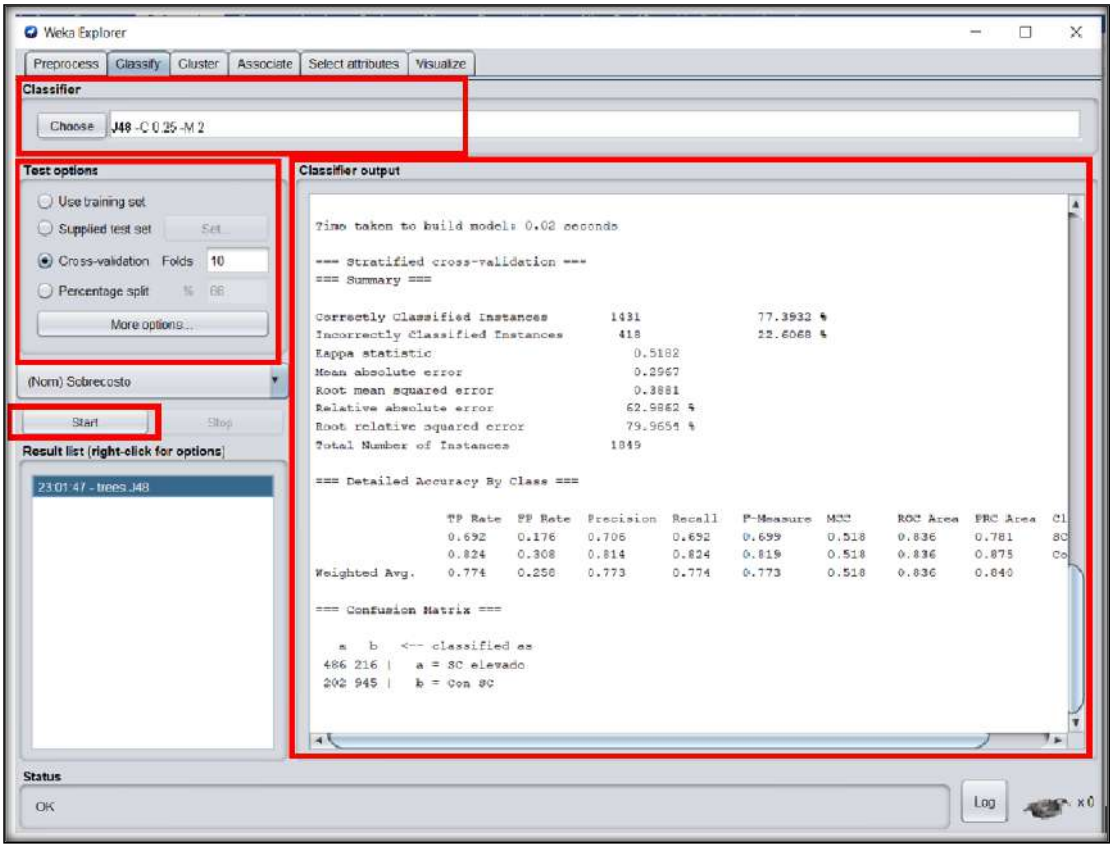


Fuente: Elaboración propia

Existen distintas formas de obtener los datos desde un fichero: Open file, Open URL y Open DB, al cargar el fichero se mostrarán sus atributos y los elementos de estos en la parte de “Selected atribute”, con una opción en la parte inferior para remover en caso sea necesario, por último, se muestra un gráfico de barras del cual se puede obtener información relevante para el trabajo.

En cuanto a la pestaña de Classify, se lleva a cabo la selección de la técnica, entre muchas, a utilizar y aplicar al fichero. “Modo de clasificación, conocida en algunas ocasiones como aprendizaje supervisado, haciendo uso de técnicas de clasificación y regresión” (Navas, 2016).

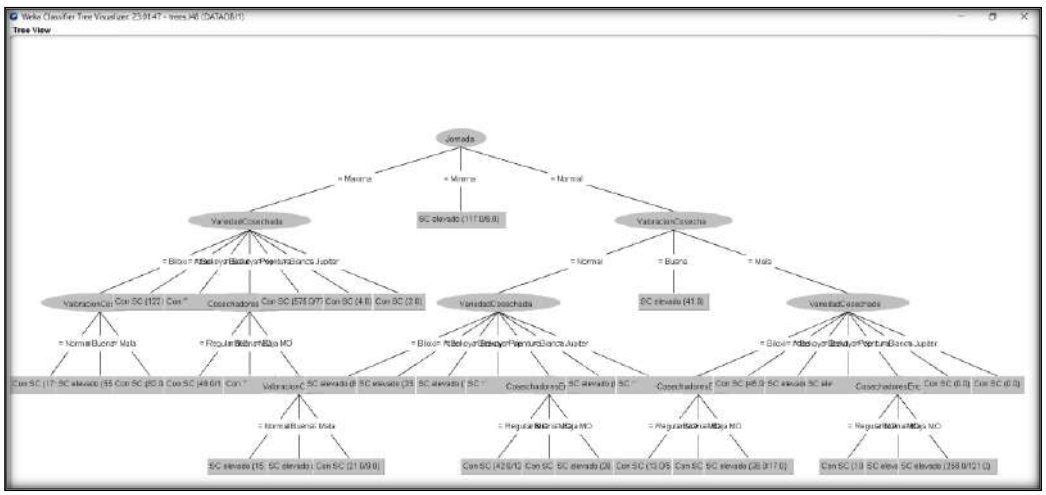
Imagen 52: Interfaz y funciones de Classify de Weka Explorer



Fuente: Elaboración propia

Comenzando por seleccionar la técnica para aplicar (Árbol de decisión J48 en el ejemplo), las opciones de la prueba en donde muestra distintas opciones a escoger se proceden a dar a Start para mostrar los resultados en la parte derecha, además al dar clic al resultado se podrá abrir una nueva ventana en donde mostrará la vista del árbol de decisión.

Imagen 53: Vista del árbol de decisión J48 de ejemplo con 5 atributos en Weka Explorer

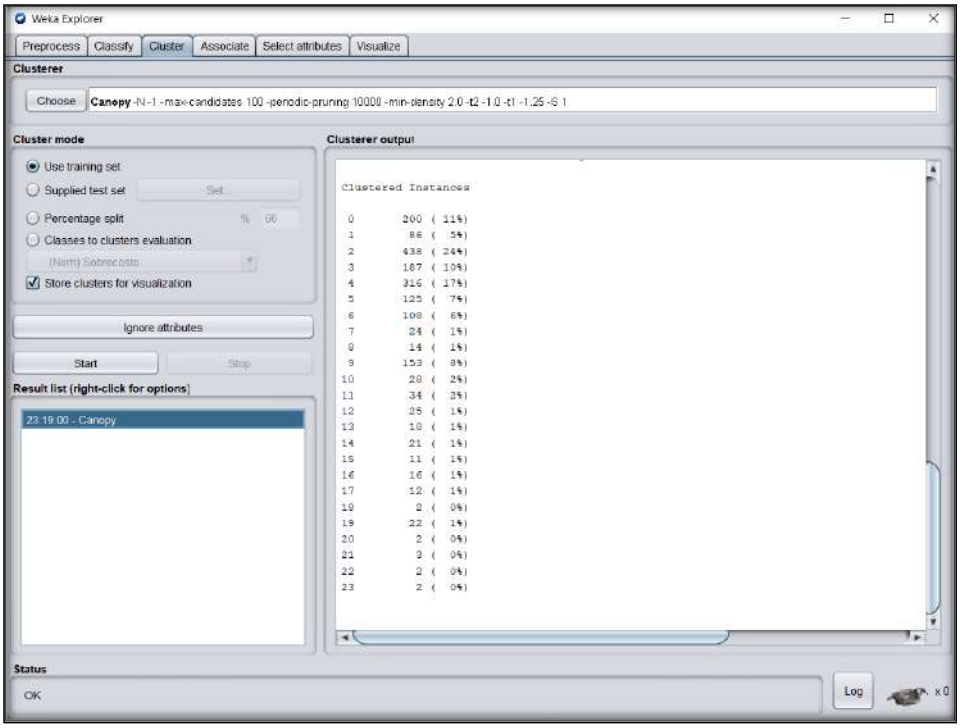


Fuente: Elaboración propia

El clasificador J48 del árbol de decisión es el más utilizado debido a su nivel de confianza superior a otros, siendo esta su característica más importante para la selección.

La tercera pestaña de la opción WEKA Explorer llamada Cluster o agrupamiento según (Navas, 2016) “permite realizar agrupamientos de los datos basándose en semejanzas y diferencias existentes entre los datos”.

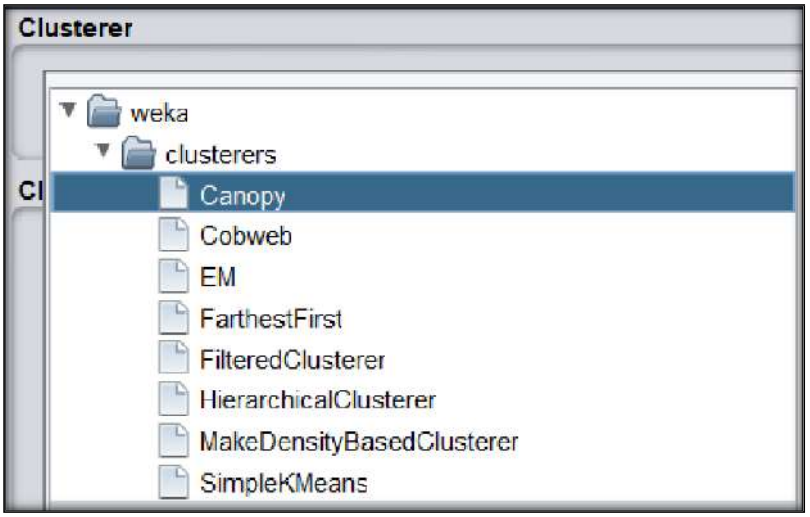
Imagen 54: Interfaz de la pestaña Cluster de Weka Explorer



Fuente: Elaboración propia

La interfaz y las funciones son muy parecidas a la pestaña de Classify, a diferencia que en Cluster se mostrarán algoritmos de agrupamiento los cuales se muestran en la siguiente imagen.

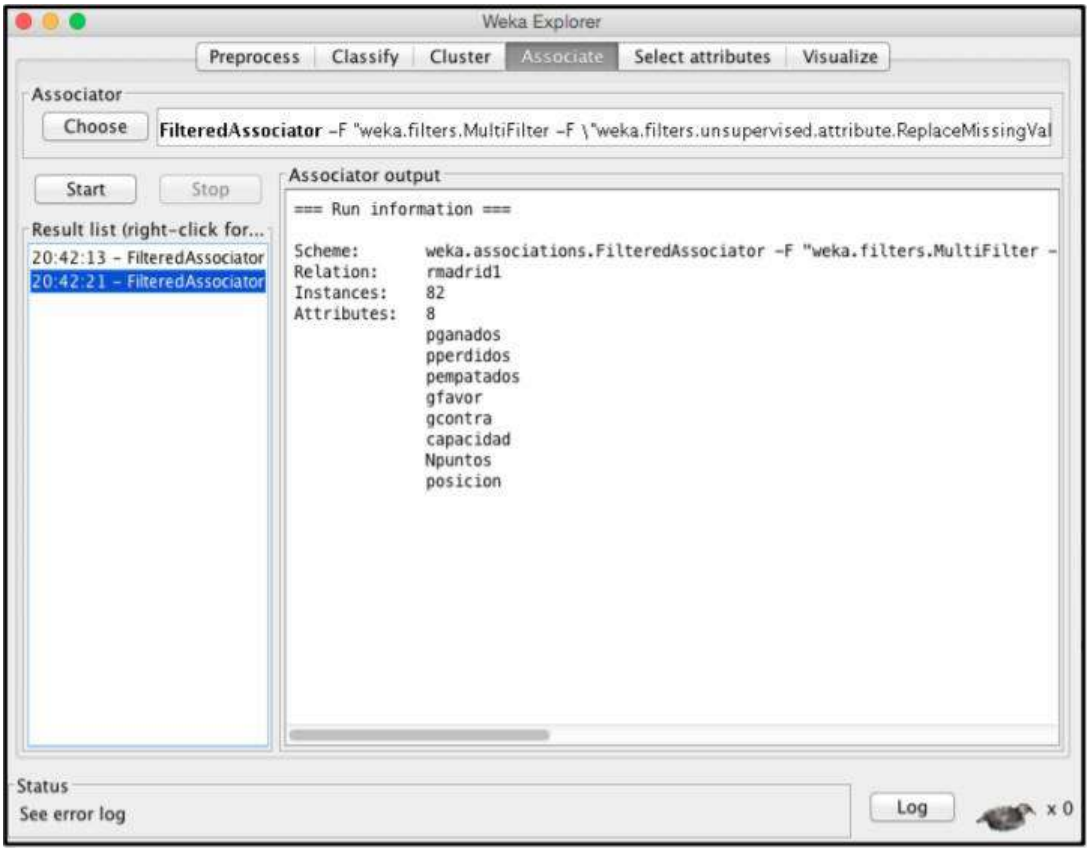
Imagen 55: Lista de algoritmos de agrupamiento



Fuente: Elaboración propia

Tal como señala el autor (Navas, 2016) “la cuarta pestaña denominada Associate se aplican métodos y algoritmos para buscar asociaciones en los datos, se selecciona el método, se configura y se presiona Start. Es necesario para discretizar los datos, preprocesar los atributos, en caso contrario el programa no se ejecutará correctamente”.

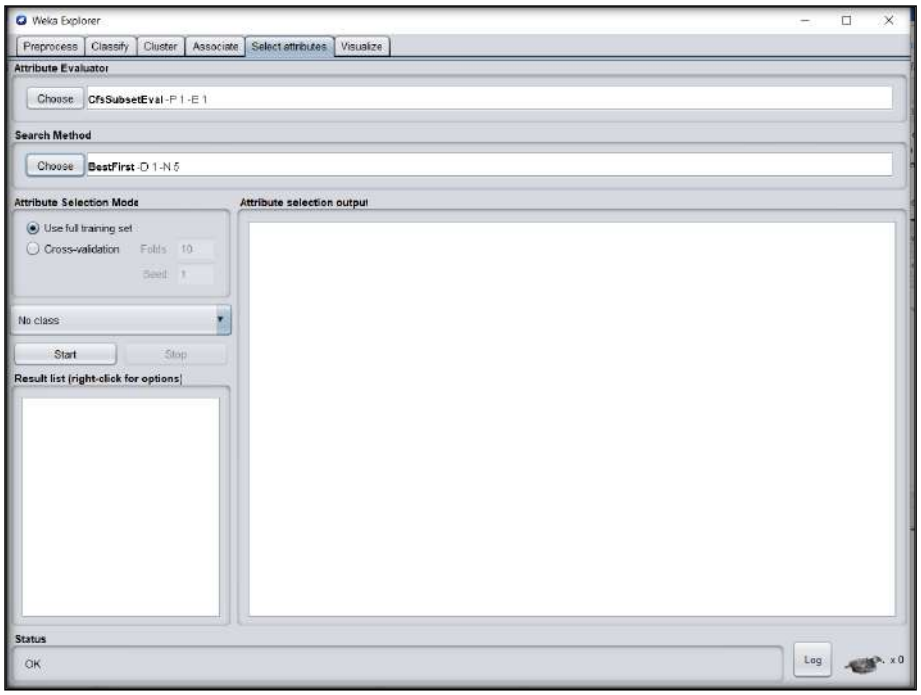
Imagen 56: Resultados del algoritmo de asociación



Fuente: Elaboración propia

La quinta pestaña corresponde a Select attributes, la cual tiene un estilo similar a las pestañas ya vistas, desde seleccionar un tipo de evaluador de atributo, un método, el modo y por último se mostrarán los resultados en la parte derecha de la interfaz.

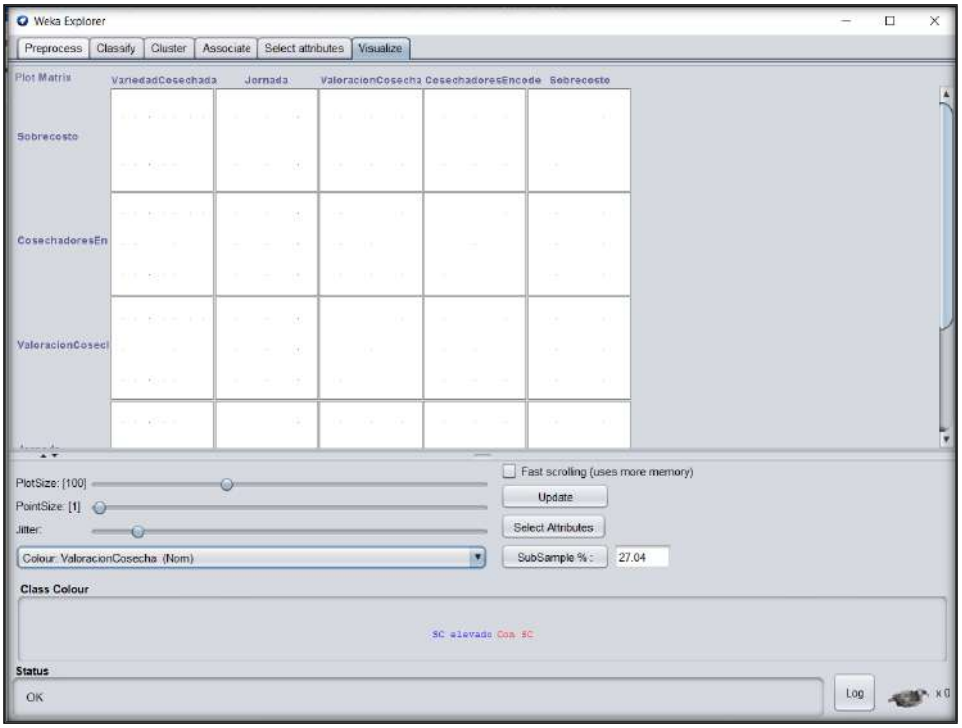
Imagen 57: Interfaz de la pestaña Select attributes de Weka Explorer



Fuente: Elaboración propia

Y por último la pestaña de Visualize, que muestra la relación de los datos de manera visual y gráficos.

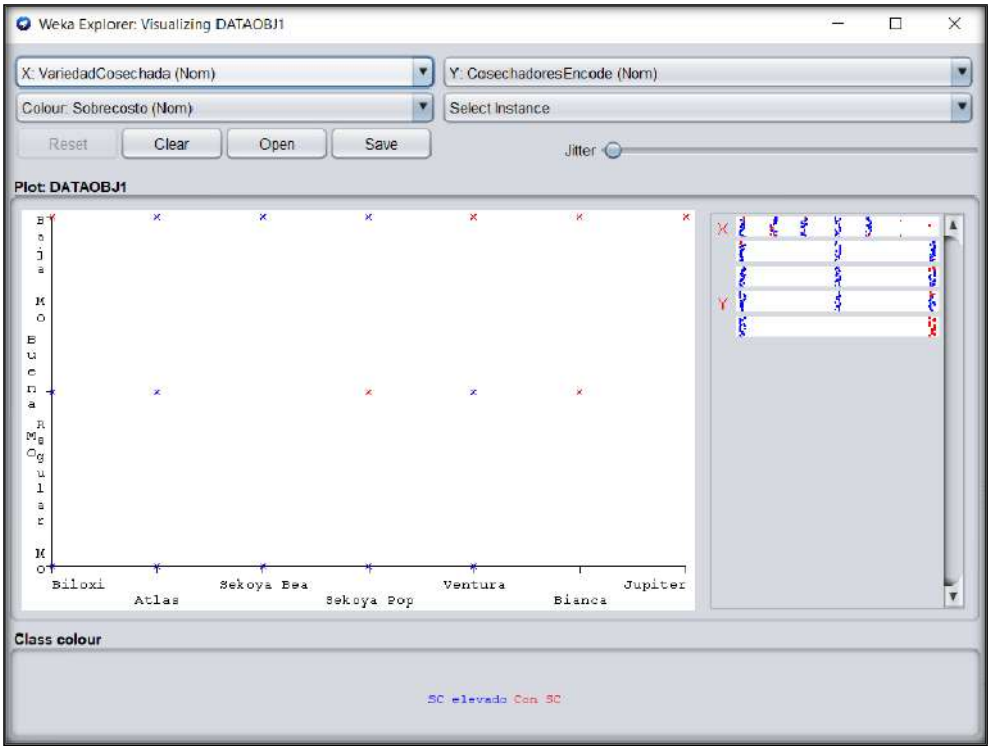
Imagen 58: Interfaz de la pestaña Visualize de Weka Explorer



Fuente: Elaboración propia

En esta pestaña se seleccionan los atributos a relacionar para mostrarse de la siguiente manera:

Imagen 59: Configuración para visualizar las relaciones en Visualize de Weka Explorer

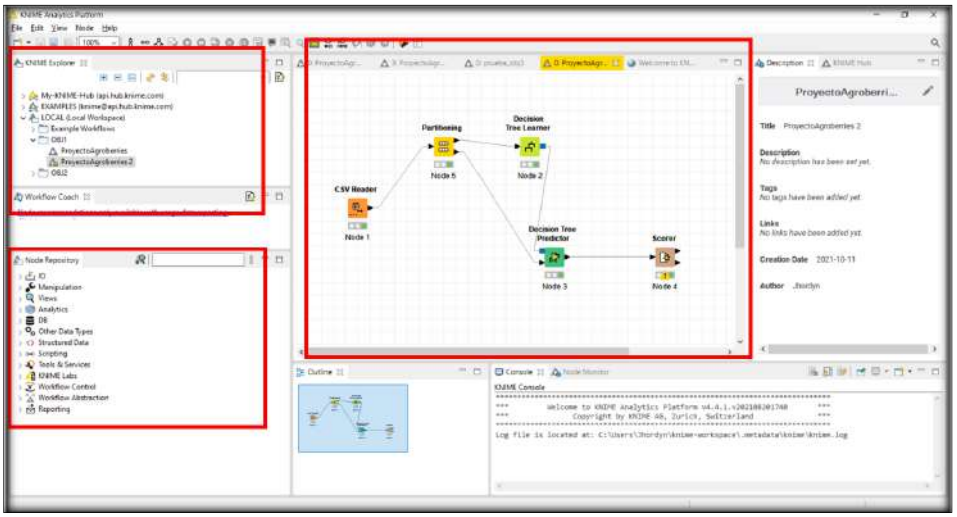


Fuente: Elaboración propia

2.3.7. SOFTWARE KNIME

La plataforma KNIME enfocada a la MD permite el desarrollo de modelos en un entorno visual. Para el trabajo se utilizó la versión 4.4.1 y la interfaz se muestra como en la siguiente imagen.

Imagen 60: Interfaz Knime Analytics Plataform



Fuente: Elaboración propia

La interfaz tiene un diseño sencillo para entenderla, comenzando por el explorador de los archivos en la parte superior izquierda, el repositorio de

nodos utilizados para la construcción de modelos en la parte inferior izquierda y el área de trabajo donde se llevará a cabo el modelo, parte del medio.

A continuación, se describe en la siguiente tabla algunos de los nodos según (Tapia, 2015).

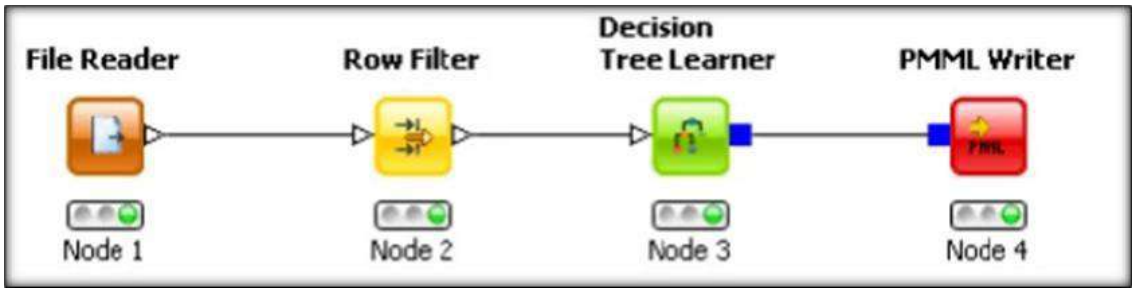
Tabla 18: Tipo de nodos que proporciona Knime

Entrada de datos	[IO > Read]
Salida de datos	[IO > Write]
Pre-procesamiento	[Data Manipulation] para normalizar, filtrar, discretizar, seleccionar variables.
Minería de datos	[Mining], para construir modelos (reglas de asociación, clustering, clasificación…)
Salida de resultados	[Data Views] para mostrar resultados en pantalla (ya sea de forma textual o gráfica).

Fuente: Tapia, 2015

Para la construcción de un modelo con estos distintos tipos de nodos se puede llevar a cabo como se muestra en la siguiente imagen.

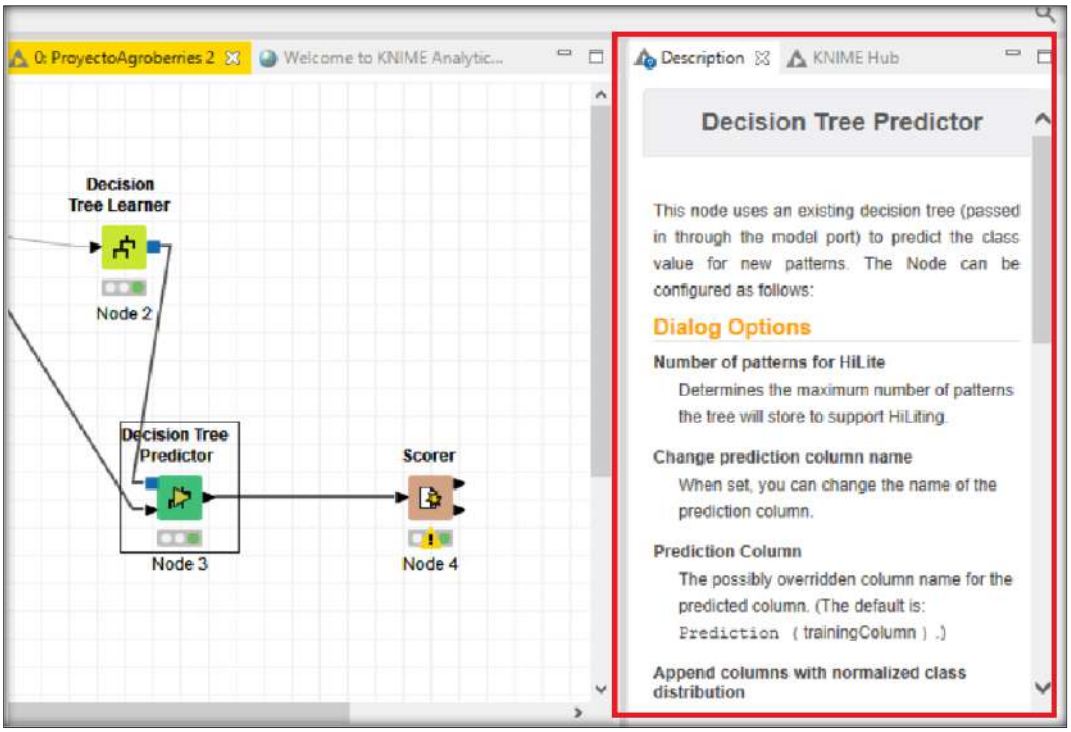
Imagen 61: Ejemplo flujo de ejecución en Knime



Fuente: Tapia, 2015

La descripción de cada nodo a usar se puede visualizar en la parte derecha del software. Sirve de ayuda para implementar correctamente cada uno de los nodos.

Imagen 62: Descripción de los nodos en software Knime

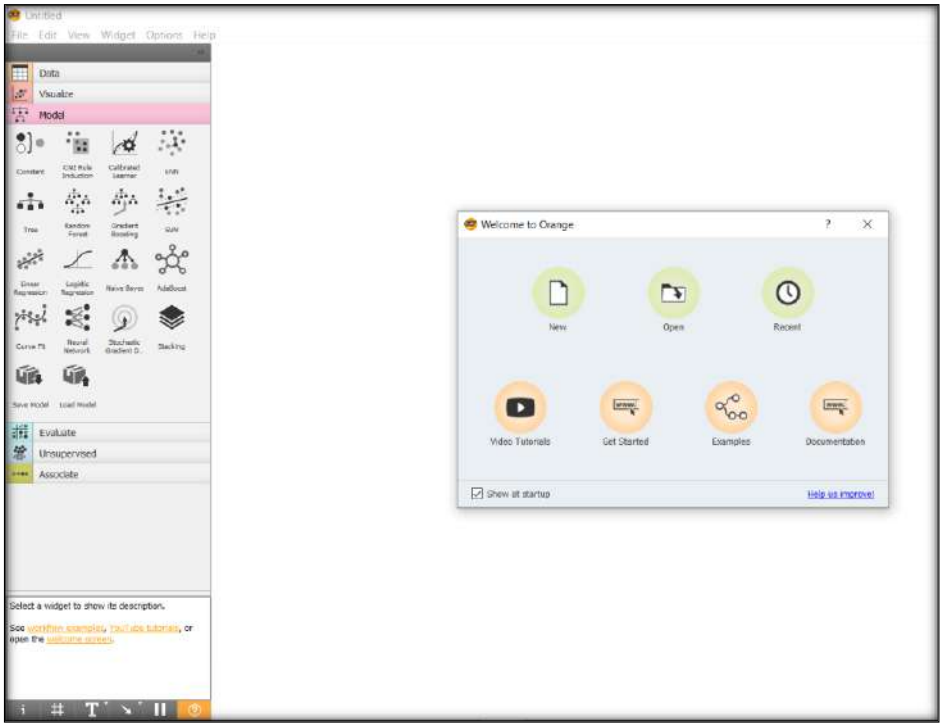


Fuente: Elaboración propia

2.3.8. SOFTWARE ORANGE

El autor (Haro, 2017) define el software de la siguiente manera: “Orange implementa algoritmos de minería de datos, operaciones de procesamiento y también ofrece visualización de datos por medio de gráficos, es un software libre y su principal función es analizar grandes cantidades de base de datos de manera automática, transformando los datos en información y optimizar tomas de decisiones.”

Imagen 63: Interfaz inicial de Orange



Fuente: Elaboración propia

Además, las funcionalidades y ventajas según (Haro, 2017) se pueden apreciar en la siguiente tabla:

Tabla 19: Funcionalidades y ventajas de Orange

Funcionalidades y ventajas de Orange	
Funcionalidades	**Ventajas**
Creación de flujos de trabajo interactivos	Visualización de información en distintos gráficos
Analiza y visualiza datos con mayor amplitud	Mayor claridad en los resultados
Rediseñar y adaptar herramientas a las necesidades del usuario y/o de la empresa	Instalación gratuita y sencilla

Fuente: Haro, 2017

El software ofrece distintos componentes para el trabajo mostrados en la siguiente imagen.

Imagen 64: Componentes de Orange

Fuente: Elaboración propia

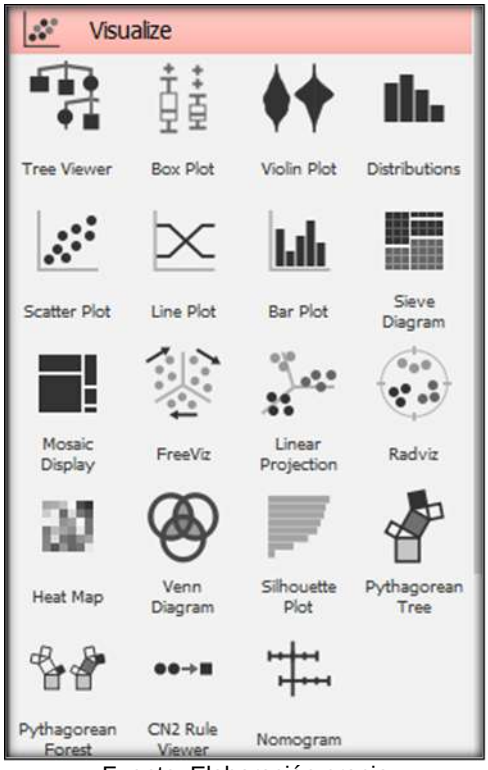
Componente Data. Permite la entrada y salida de datos.

Imagen 65: Componente Data de Orange

Fuente: Elaboración propia

Componente Visualize. Permite la visualización de datos.

Imagen 66: Componente Visualize en Orange



Fuente: Elaboración propia

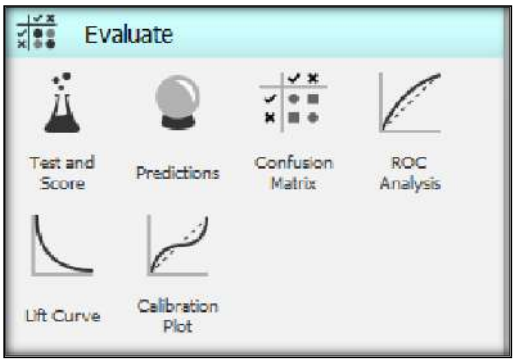
Componente Model. Permite realizar regresión y clasificación.

Imagen 67: Componente Model en Orange

Fuente: Elaboración propia

Componente Evaluate. Permite realizar evaluaciones al conjunto de datos.

Imagen 68: Componente Evaluate en Orange



Fuente: Elaboración propia

Componente Unsupervised. Proporciona algoritmos no supervisados, es decir, algoritmos de agrupamiento.

Imagen 69: Componente unsupervised en Orange

Fuente: Elaboración propia

Componente Associate. Proporciona reglas de asociación y conjuntos de datos frecuentes.

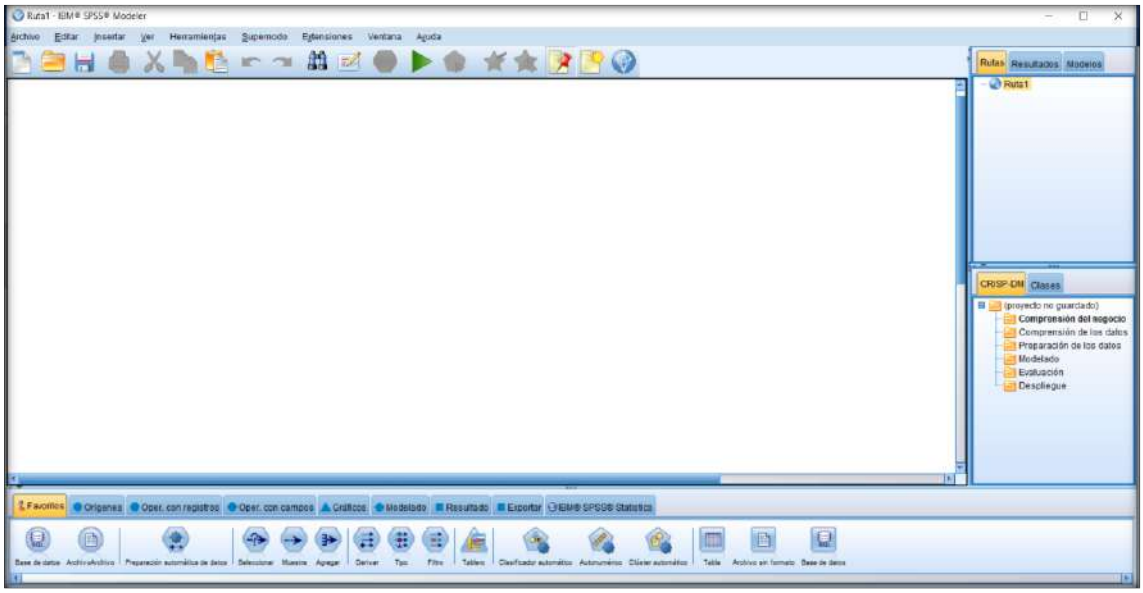
Imagen 70: Componente Associate en Orange

Fuente: Elaboración propia

2.3.9. SOFTWARE SPSS MODELER

El software SPSS MODELER de IBM es utilizado para análisis de texto y MD, en ella se construyen modelos predictivos y analíticos, entre otras. El trabajo hace uso de este software para aplicar reglas de asociación.

Imagen 71: Interfaz de IBM SPSS Modeler



Fuente: Elaboración propia

En la parte inferior se puede observar las opciones disponibles para llevar a cabo el modelo. Como en otras aplicaciones, se comienza desde un origen de datos en dónde se puede seleccionar desde una base de datos, archivos SAS, Excel, XML, etc.

Imagen 72: Pestaña Orígenes de IBM SPSS Modeler

Fuente: Elaboración propia

En cuanto al modelado, se puede seleccionar entre distintas técnicas de analytic server, clasificación, asociación y segmentación.

Imagen 73: Pestaña Modelado de IBM SPSS Modeler

Fuente: Elaboración propia

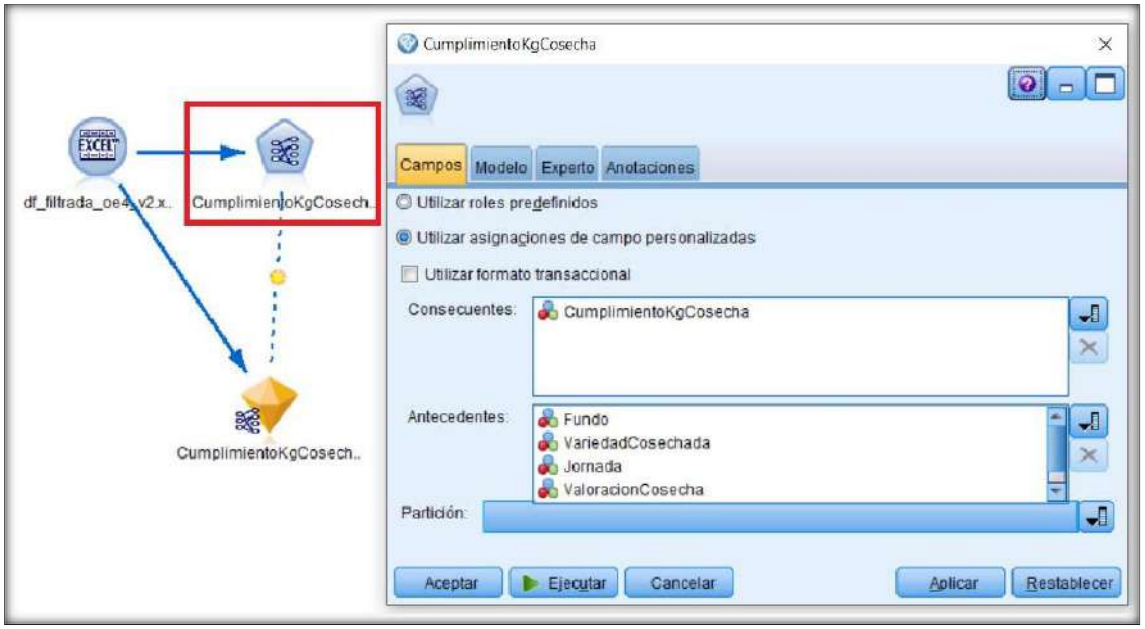
La configuración de origen del modelo se puede observar en la siguiente imagen, en dónde desde la pestaña origen de un archivo de Excel se importa el archivo y muestra los filtros, tipos de los datos y anotaciones.

Imagen 74: Vista del origen en IBM SPSS Modeler

Fuente: Elaboración propia

La configuración de la regla de asociación en este caso, APRIORI, se muestra en la imagen en donde se lleva a cabo cada uno de los puntos necesarios para su correcta ejecución.

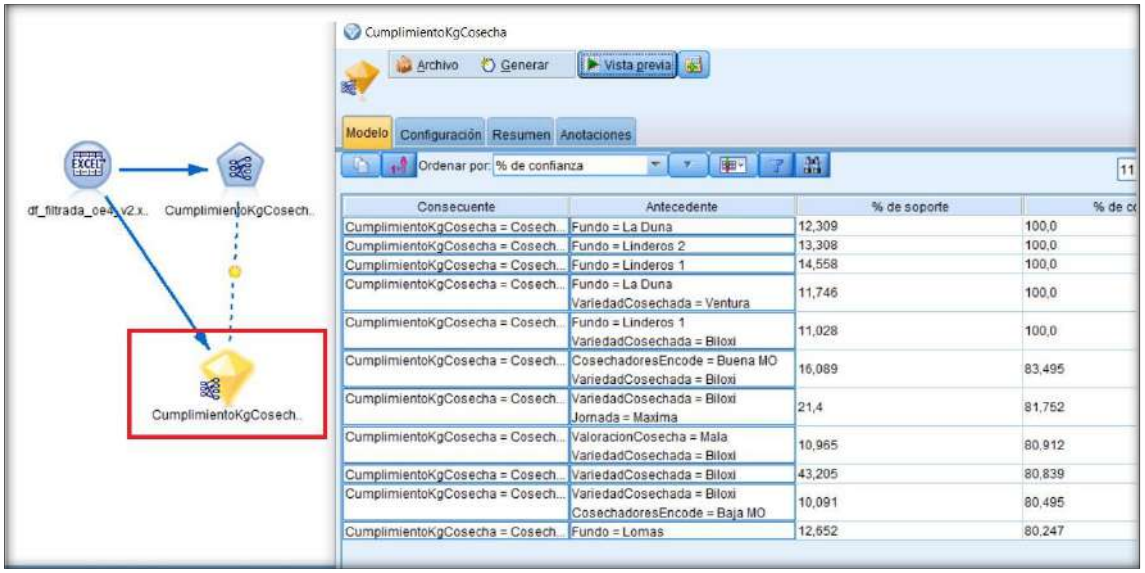
Imagen 75: Vista del Modelado (asociación, A PRIORI) en IBM SPSS Modeler



Fuente: Elaboración propia

El resultado se genera automáticamente en la parte inferior y muestra la vista de los resultados de acuerdo con las configuraciones realizadas anteriormente.

Imagen 76: Vista de los resultados en IBM SPSS Modeler



Fuente: Elaboración propia

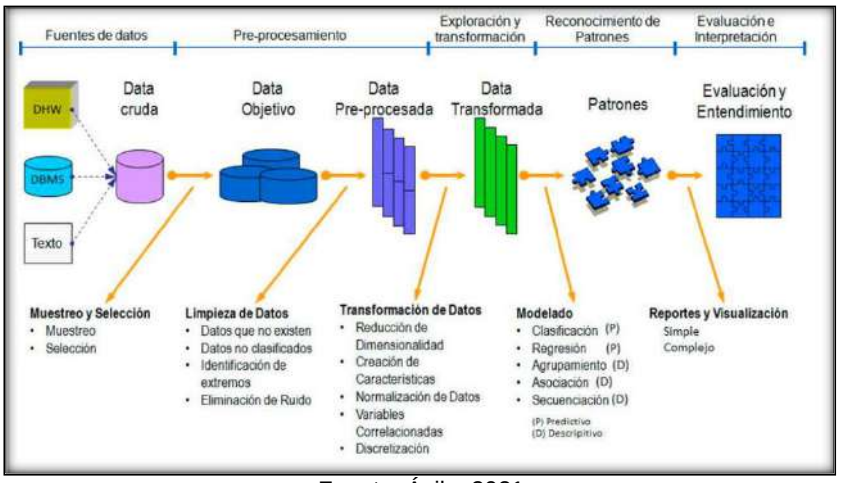
2.3.10. PREPROCESAMIENTO EN PYTHON

PYTHON

Según (Mamgain, 2018): “Python es popular en el análisis de datos porque su sintaxis es fácil de entender y legible. Opera gran conjunto de datos conocido como Big Data por lo que es muy usado por científicos de datos. Además, Python tiene bibliotecas para visualizar análisis estadísticos y procesamiento de lenguaje natural, esta inmensa biblioteca proporciona a los científicos de datos gran cantidad de funcionalidades. La ventaja de Python es su interacción directa con el código o usando la terminal u otras herramientas como cuadernos. La minería de datos y el aprendizaje automático son un subconjunto de la inteligencia artificial y ambos son procesos relacionados en los que se analizan los datos, por lo que es fundamental contar con herramientas que permitan una rápida iteración y una fácil interacción (…)”.

Gracias al lenguaje de programación Python, utilizado en este proyecto para el preprocesamiento de datos y llevar a cabo la minería de datos, tiene como función el análisis ordenado para extraer de manera automatizada la inteligencia que estos datos poseen. El trabajo que se puede llevar a cabo en Python se puede visualizar en la siguiente imagen.

Imagen 77: Análisis de datos en Python



Fuente: Ávila, 2021

- **Muestreo y Selección.** Se identifican y seleccionan variables relevantes de los datos. A partir de la muestra se estiman características como media, total, proporción, etc.
- **Exploración.** Aplicar técnicas de análisis exploratorio para identificar valores inusuales, extremos, discontinuidades, u otra peculiaridad de estos.
- **Limpieza.** Solventar inconvenientes debido a valores atípicos, faltantes y/o erróneos.

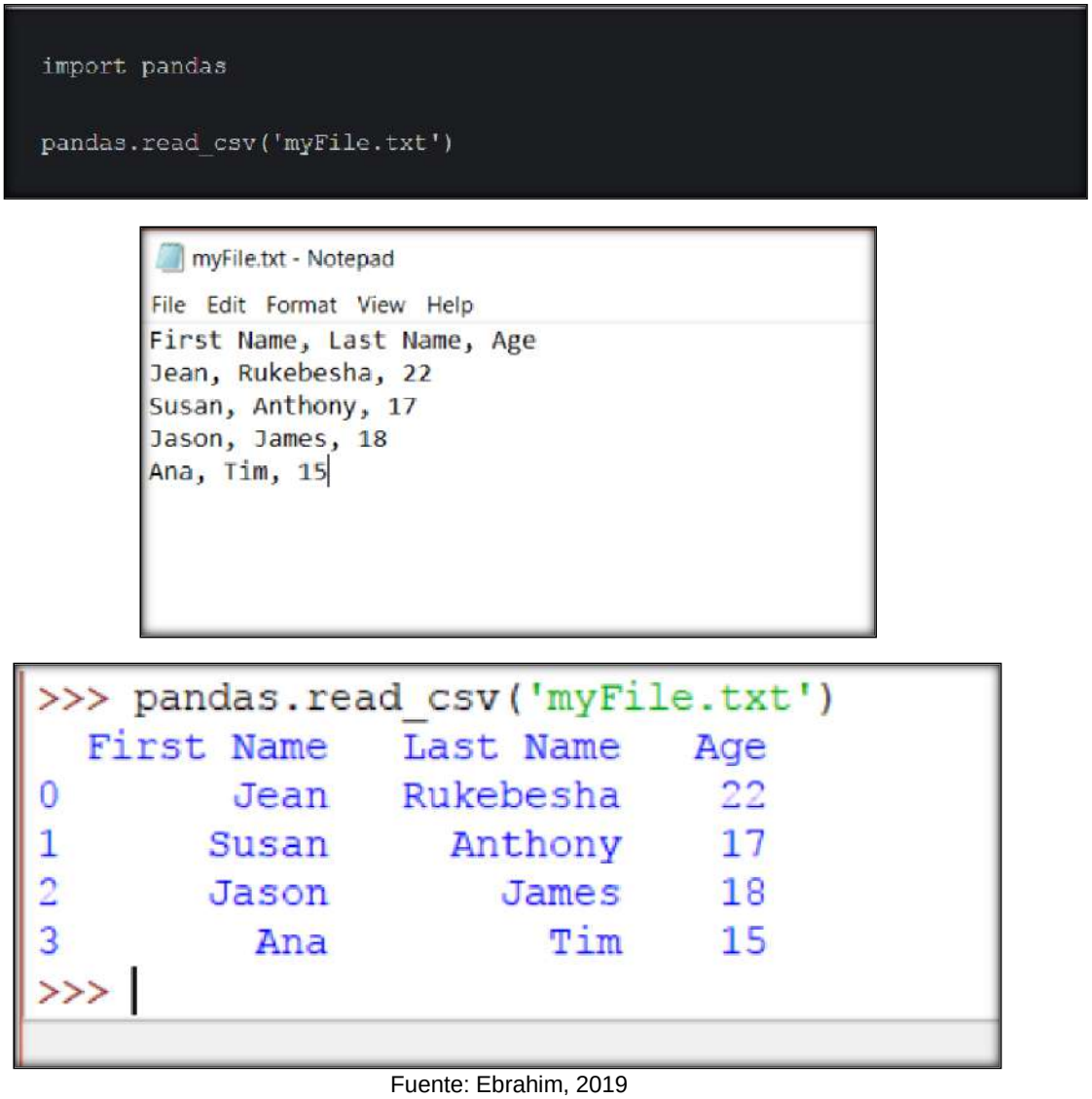
- **Transformación.** Llevar a cabo una normalización o una estandarización mediante técnicas de reducción o aumento de dimensiones.

LIBRERÍA PANDAS

Python hace uso de librerías con distintas funciones, las cuales pueden ser llamadas por código. En este caso la librería Pandas es la encargada de proporcionar análisis y manipulación de datos.

"Es una biblioteca muy prometedora de representación de datos, filtrado y programación estadística. La pieza más importante en pandas es el DataFrame donde almacena y juega con los datos." (Ebrahim, 2019).

Imagen 78: Proceso leer archivo de texto con librería Pandas



Fuente: Ebrahim, 2019

2.4. SISTEMA DE HIPÓTESIS

2.4.1. HIPÓTESIS

La aplicación de la minería de datos y aprendizaje automático sirve como herramienta de soporte para la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022.

2.4.2. VARIABLES E INDICADORES

- **Variables:**
  - **V. independiente:** Sistema de minería de datos y aprendizaje automático.
  - **V. dependiente:** Toma de decisiones.

- **Indicadores:**
  - **Sistema de minería de datos y aprendizaje automático (V. independiente):** confiabilidad, usabilidad.
  - **Toma de decisiones (V. dependiente):** eficacia, eficiencia, grado de satisfacción.

Tabla 20: Cuadro de operacionalización de variables.

Variables	Definición conceptual	Indicadores	Tipo	Técnica	Instrumento
Sistema de minería de datos y aprendizaje automático	Se define en muchos casos como un proceso que identifica patrones de interés en un conjunto de datos que son usados en la toma de decisiones.	% Precisión de las distintas técnicas de minería de datos y aprendizaje automático usadas para la resolución de los problemas planteados que a su vez sean muestren ser fiables y usables.	Cuantitativo	Observación	▪ Guía de observación
Toma de decisiones	Es el proceso para la elección de una alternativa frente a muchas otras, mayormente se basa en diversos y variados motivos.	% Grado de eficacia, eficiencia y satisfacción de los tomadores antes y después de realizada la minería de datos.	Cuantitativo	Encuesta	▪ Cuestionario ▪ Escala

CAPÍTULO III:
METODOLOGÍA EMPLEADA

CAPÍTULO III: METODOLOGÍA EMPLEADA

3.1. TIPO Y NIVEL DE INVESTIGACIÓN

Tipo de investigación: experimental. Nivel de investigación: explicativo.

3.2. POBLACIÓN Y MUESTRA DE ESTUDIO

- **Población:**
  Los datos de producción y transporte de arándanos de Agroberries S.A.C.

- **Muestra:**
  Los datos de producción y transporte de arándanos de Agroberries S.A.C. del año 2020 de la semana 29.

3.3. DISEÑO DE INVESTIGACIÓN

Es una investigación cuantitativa porque se expresará de manera numérica; es una investigación descriptiva puesto que describe la satisfacción de los tomadores de decisiones con respecto a la minería de datos realizada; es una investigación retrospectiva ya que tomará datos del año 2020, en específico de la semana 29 de ese año. Por último, la investigación es de carácter estadístico descriptivo debido a que se usará la recolección, el almacenamiento, ordenamiento y se harán gráficos y tablas.

GE: O1 …… X O2

G.E: Grupo experimental.

O1: Pre-Test.

O2: Post-Test.

X: Manipulación de la variable independiente.

3.4. TÉCNICAS E INSTRUMENTOS DE INVESTIGACIÓN

- **Técnica de observación:** Se hará uso de la técnica para la interpretación de los resultados obtenidos tras haber encontrado las relaciones de los datos con diversas técnicas que conciernen a Machine Learning y, claro, minería de datos.
- **Instrumento de guía de observación:** Es un instrumento que recuerda al investigador los puntos clave por los que se realiza la observación, así también los temas de interés asociados a cada uno.
- **Técnica de encuesta:** Para hacer uso de la técnica de encuesta, se aplicará a los trabajadores disponibles de la empresa Agroberries S.A.C. para medir la toma de decisiones y el evento que generan las mismas, de tal modo que podamos evaluar lo positivo o negativo desde la perspectiva operaria.
- **Instrumento de cuestionario:** Se trata de un instrumento que reúne un conjunto de preguntas estandarizadas, que se denominan ítems, y estos siguen un esquema fijo para la recolección de datos de forma individual sobre uno o varios temas en específico.
- **Instrumento de escala:** El instrumento de escala hace referencia al procedimiento de asignación de una serie de objetos a números según reglas previamente especificadas; en síntesis, se trata de ubicar los objetos medidos en una secuencia continua de números.

3.5. PROCESAMIENTO Y ANÁLISIS DE DATOS

a) CUADRO O TABLAS ESTADÍSTICAS

Necesario para facilitar la visualización de los datos en filas y columnas que se presentarán de manera ordenada gracias a ello; se usará para acomodar la base de datos inicial y presentar información más detallada, resultante del proceso y análisis estadístico.

b) GRÁFICOS ESTADÍSTICOS

Para una comprensión de los datos de manera sencilla, rápida y global se presentan visualmente en gráficos estadísticos; se usará la hoja de cálculo de Excel para la presentación de los datos agrupados de acuerdo con nuestras necesidades, el tipo de gráfico más usado fue el histograma.

c) MEDIDAS ESTADÍSTICAS

El procesamiento y análisis incluye las siguientes medidas estadísticas:

i. **MEDIA ARITMÉTICA:** “Es el valor obtenido de la suma de todos los datos dividido entre el número de datos sumados, obteniendo el promedio” (Ruiz, 2020).

Datos desagrupados:

$$\bar{x} = \sum xi/N$$

Donde:

$\bar{x}$: Media aritmética

$\sum$: Sumatoria

xi: Datos

N: Total de datos

Datos agrupados:

$$\bar{x} = fi * xi \ / \ N$$

Donde:

$\bar{x}$: Media aritmética

fi: Frecuencia

xi: Marca de clase (Suma de L.I. + L.S. entre 2)

N: Total de datos

ii. **DESVIACIÓN ESTÁNDAR:** “De gran utilidad para la estadística descriptiva, se define como la raíz cuadrada de la varianza, informa la media de distancias que tienen los datos respecto a su media aritmética” (Quintana, 2017).

$$S = \sqrt{\sum ni(xi - \bar{x})^2/N - 1}$$

Donde:

S: Desviación estándar
xi: Valores individuales
ni: Frecuencia de valor X
N: Casos

iii. **T-STUDENT:** “Permite decidir si dos variables aleatorias normales y con la misma varianza tienen medias diferentes (…). Opera diciendo si una diferencia en la media muestral entre dos muestras es estadísticamente significativa, y poder afirmar que las dos muestras corresponden a distribuciones de probabilidad de media poblacional distinta, o afirmar que la diferencia de medias puede deberse a oscilaciones estadísticas al azar” (Fátima & Serrano, 2015).

$$Prueba\ T\ de\ muestras\ independientes = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Donde:

$$\boldsymbol{S^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_j - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

CAPÍTULO IV:
PRESENTACIÓN DE RESULTADOS

CAPÍTULO IV: PRESENTACIÓN DE RESULTADOS

A continuación, se desarrolla cada uno de los objetivos propuestos en la investigación:

4.1. PROPUESTA DE INVESTIGACIÓN

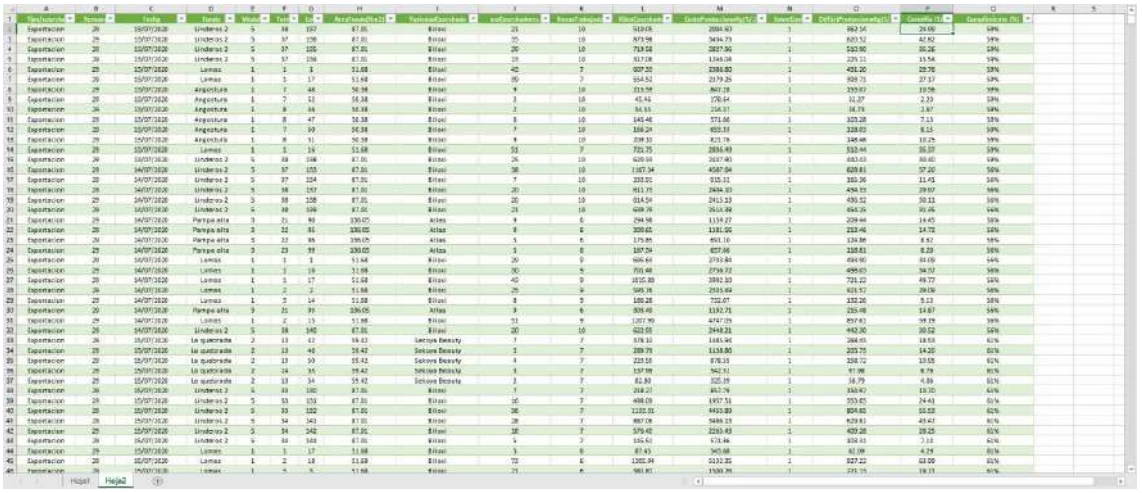
Nuestra propuesta de investigación es la aplicación de la minería de datos y aprendizaje automático a través del uso de herramientas de software como Weka, Orange, SPSS Modeler y KNIME con el fin de obtener resultados relevantes que se obtienen de los datos procesados con el lenguaje de programación Python y que sean una muestra fehaciente de qué la minería de datos sirve para la mejora en la toma de decisiones, una mejor administración del tiempo y que puede ayudar con suma relevancia en los procesos de transporte y producción.

4.2. ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

4.2.1. RESULTADO 1 vs OE1: ANALIZAR LA DATA DEL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS DE LA EMPRESA AGROBERRIES PARA FORMAR UN DATASET COMPUESTO SÓLO POR AQUELLAS CARACTERÍSTICAS A USAR.

Recibimos la data del área de producción y transporte de la empresa en un Excel el cual representaba un informe con tablas dinámicas por lo cual se tenía que procesar esos datos en uno solo para trabajarlo como un solo dataset.

Imagen 79: Dataset consolidado con características a usar



Fuente: Elaboración propia

El dataset presentado consta de los siguientes atributos:

- Tipo de cosecha (exportación o nacional).
- Semana (temporada).
- Fecha de la cosecha.
- El detalle de la tierra en dónde se cosechó (fundo, módulo, turno y lote).
- El área de la tierra en dónde se cosechó.
- La variedad del arándano cosechado.
- El número de cosechadores que participaron de la cosecha.
- Las horas trabajadas por cosecha.
- Kilos cosechados.
- El costo de producción por los kilos cosechados (en soles).
- Columna sobre costo qué indica si hubo o no sobrecosto en la cosecha.
- Si hubo sobre costo, indica la cantidad que se sobre pasó en valor negativo y si no hubo sobrecosto entonces indica en cantidad positiva lo que se “ahorró”.
- El costo de la Kia (transporte) por kilos cosechados.
- El porcentaje de cumplimiento semanal de Kias.

4.2.2. RESULTADO 2 vs OE2: IMPLEMENTAR TÉCNICA DE ÁRBOL DE DECISIÓN PARA PREDECIR CON BASE EN EL SOBRECOSTO.

Para realizar la implementación del árbol de decisión se siguieron una serie de pasos mostrados a continuación:

1. Identificamos los atributos a usar, que en este caso serán – de todo el dataset – los siguientes:
   - Horas trabajadas.
   - Kilos cosechados.
   - Nro de cosechadores.
   - Variedad de cosecha.
   - Sobrecosto.

2. Tras haber sido identificadas, tuvimos que transformar estas columnas de numéricas a nominales, por tanto, quedaron de la siguiente manera:

   - **Horas trabajadas**, fue transformada a **Jornada** y se compone de 3 tipos de clases:
     - Mínima: si la jornada fue menor o igual a 7 horas.
     - Normal: si la jornada fue mayor a 7 horas y menor a 9.5 horas
     - Máxima: si la jornada fue mayor o igual a 9.5 horas.
   - **Kilos cosechados**, fue transformada a **varolacionCosecha** y se compone de 3 tipos de clases:
     - Mala: si los kilos cosechados fueron menor o igual a 250 kilos.
     - Normal: si los kilos cosechados fueron mayores a 250 y menor a 800 kilos.
     - Buena: si los kilos cosechados fueron mayores a 800.
   - **Nro de cosechadores**, fue transformada a **CosechadoresEncode** y se compone de 3 tipos de clases:
     - Baja MO: si el número de cosechadores fueron menor o igual a 8.
     - Regular MO: si el número de cosechadores fueron mayor a 8 y menor a 22.
     - Buena MO: si el número de cosechadores fue mayor o igual a 22.

     Tener en cuenta que MO se refiere a mano de obra.

   - **Variedad de cosecha**, es la única variable que no se modificó puesto que ya es nominal y de por sí, se categoriza de la siguiente manera:
     - Biloxi.

- Atlas.
- Sekoya Beauty.
- Sekoya Pop.
- Ventura.
- Bianca.
- Jupiter.

- **Sobrecosto,** esta variable es un caso en particular debido a que se tenían tablas numéricas (1 indicaba sobrecosto y 0 indicaba no sobrecosto). Entonces, se tomó solo en cuenta aquellas que tenían sobrecosto y posterior a ello se subdividió en:
  - SC elevado**.**
  - Con SC.

  Tener en cuenta que <u>SC significa sobrecosto</u>.

3. La transformación hacia variables nominales se hizo con el lenguaje de programación Python (preprocesamiento de datos).

Imagen 80: Preprocesamiento para la variable "HorasTrabajadas"

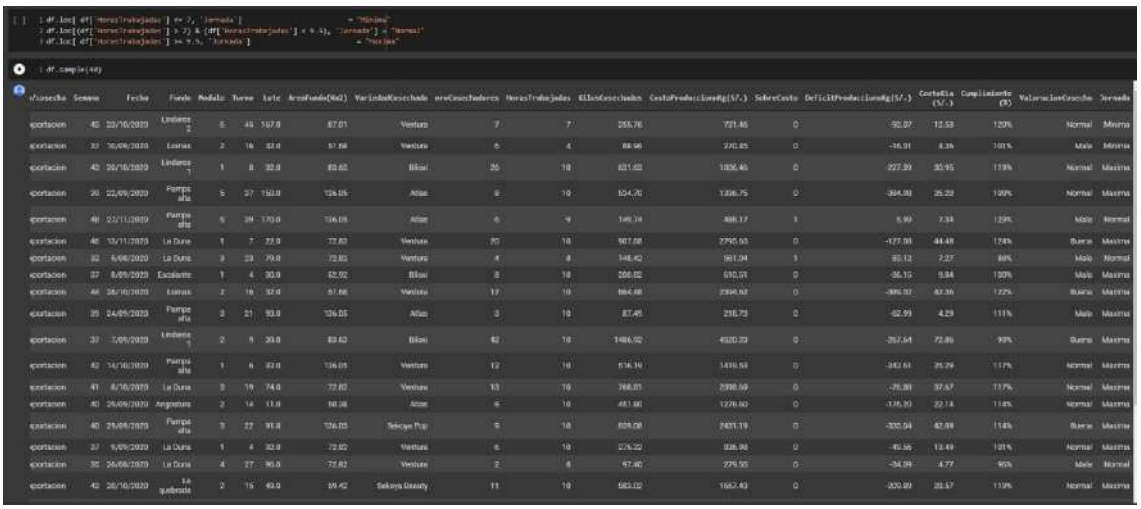


Imagen 81: Preprocesamiento para la variable “KilosCosechados”

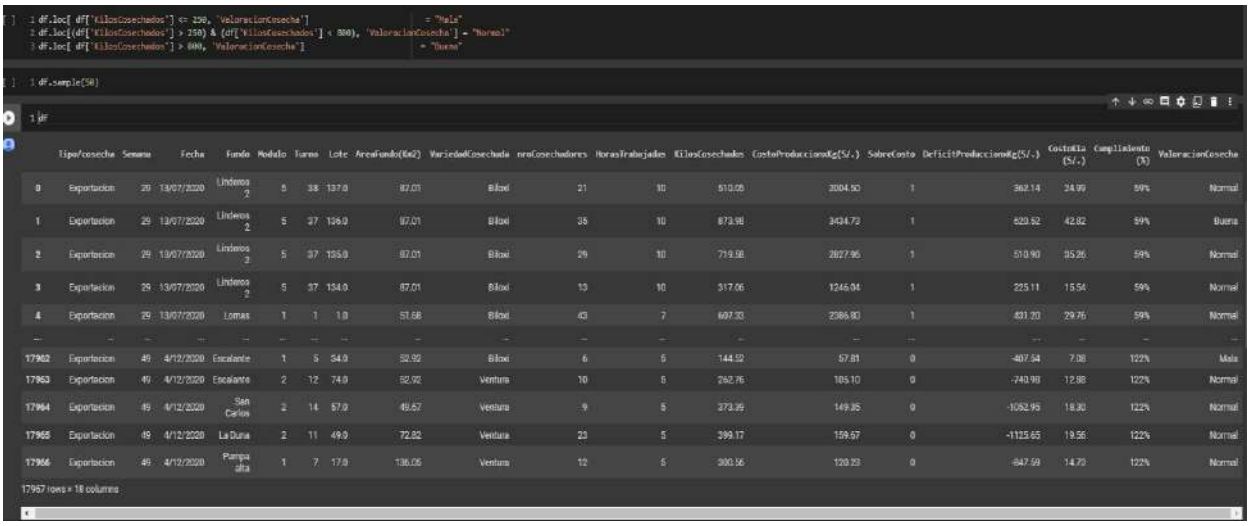


Imagen 82: Preprocesamiento para la variable “NroCosechadores”

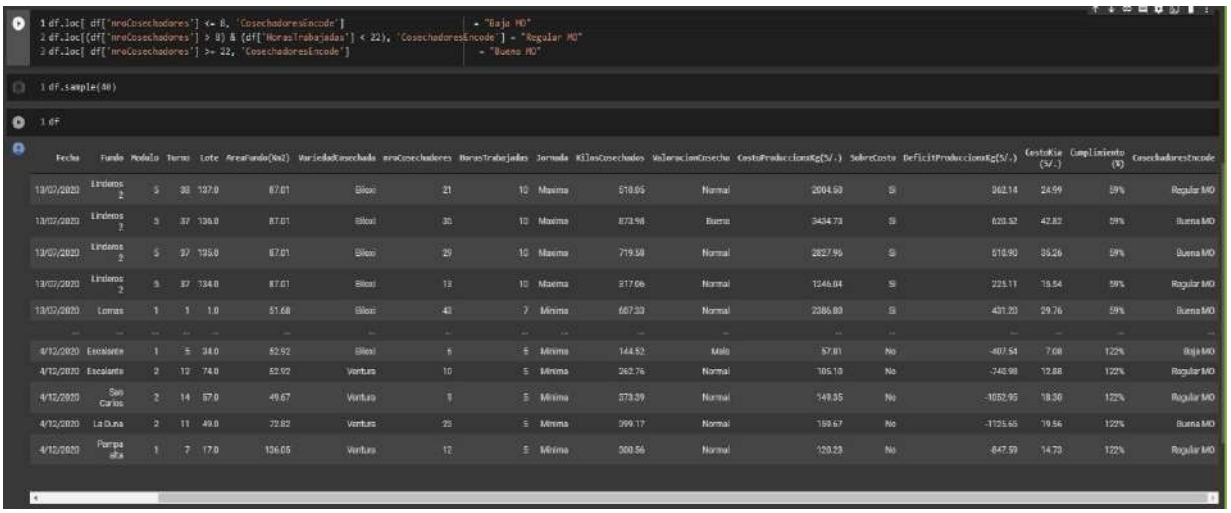
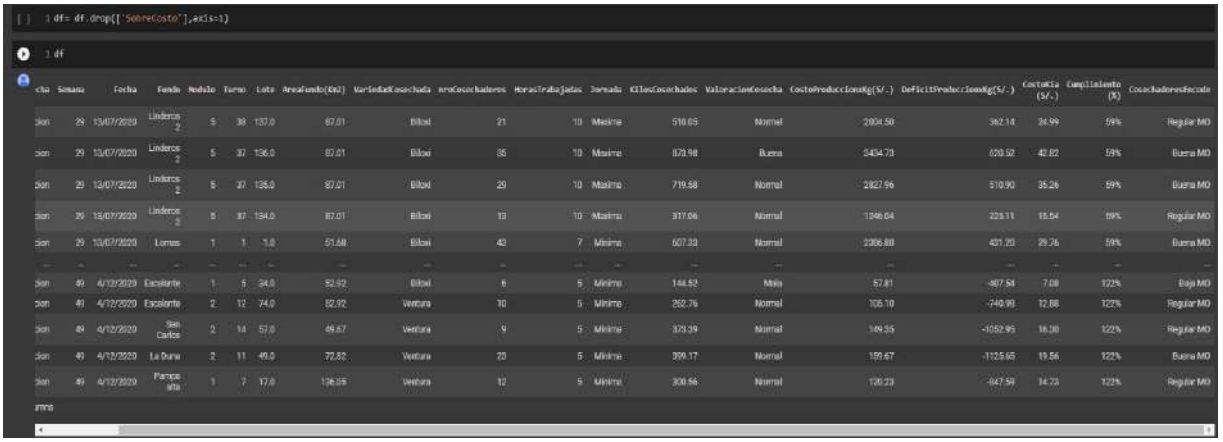
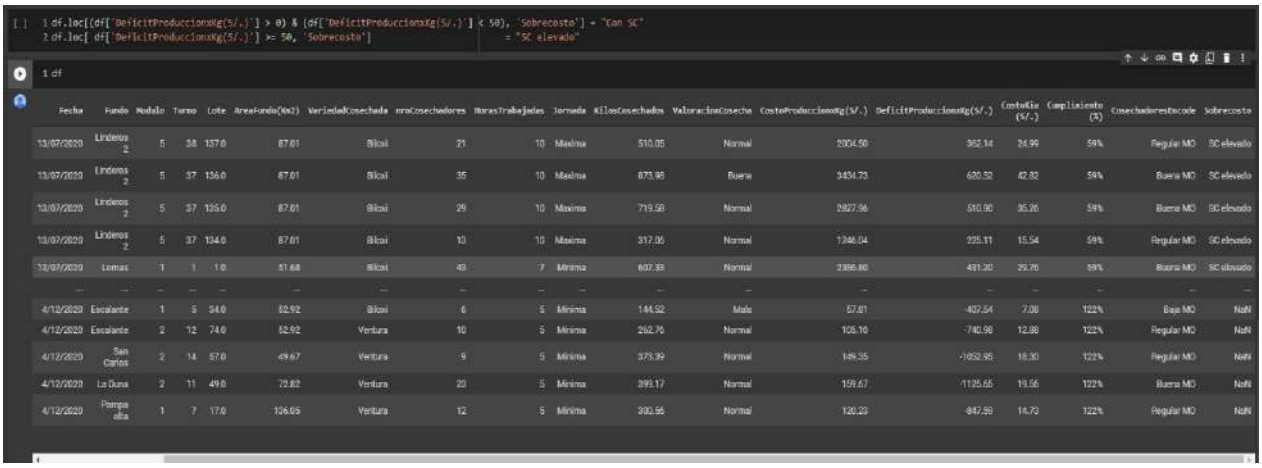


Imagen 83: Preprocesamiento para la variable “Sobrecosto”

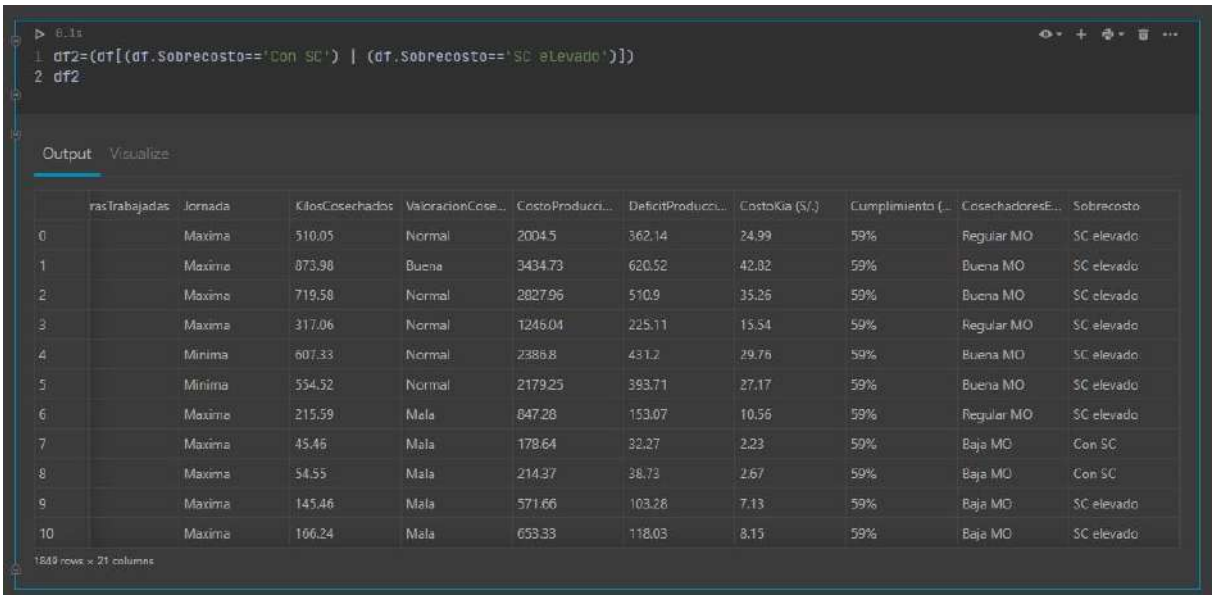




```
df.loc[(df['DeficitProduccionXKg(S/.)'] > 0) & (df['DeficitProduccionXKg(S/.)'] < 50), 'Sobrecosto'] = "Con SC"
df.loc[ df['DeficitProduccionXKg(S/.)'] >= 50, 'Sobrecosto']                          = "SC elevado"
```

```
df
```

Fecha	Fundo	Modulo	Turno	Lote	AreaFundo(Ha2)	VariedadCosechada	nroCosechadores	HorasTrabajadas	Jornada	KilosCosechados	ValoracionCosecha	CostoProduccionXKg(S/.)	DeficitProduccionXKg(S/.)	CostoKia (S/.)	Cumplimiento (%)	CosechadoresEscode	Sobrecosto
13/07/2020	Linderos 2	5	38	137.0	87.01	Bikoi	21	10	Maxima	510.05	Normal	2004.50	362.14	24.99	59%	Regular MO	SC elevado
13/07/2020	Linderos 2	5	37	136.0	87.01	Bikoi	35	10	Maxima	873.98	Buena	3434.73	620.52	42.82	59%	Buena MO	SC elevado
13/07/2020	Linderos 2	5	37	135.0	87.01	Bikoi	29	10	Maxima	719.58	Normal	2827.96	510.90	35.26	59%	Buena MO	SC elevado
13/07/2020	Linderos 2	5	37	134.0	87.01	Bikoi	10	10	Maxima	317.06	Normal	1246.04	225.11	15.54	59%	Regular MO	SC elevado
13/07/2020	Lomas	1	1	1.0	51.68	Bikoi	49	7	Minima	607.33	Normal	2386.80	431.20	29.76	59%	Buena MO	SC elevado
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4/12/2020	Escalante	1	5	94.0	52.92	Bikoi	6	5	Minima	144.52	Mala	57.81	-402.54	7.08	122%	Baja MO	NaN
4/12/2020	Escalante	2	12	74.0	52.92	Ventura	10	5	Minima	262.76	Normal	105.10	-740.98	12.88	122%	Regular MO	NaN
4/12/2020	San Carlos	2	14	57.0	49.67	Ventura	9	5	Minima	373.39	Normal	149.35	-1052.95	18.30	122%	Regular MO	NaN
4/12/2020	La Duna	2	11	49.0	72.82	Ventura	23	5	Minima	399.17	Normal	159.67	-1125.65	19.56	122%	Buena MO	NaN
4/12/2020	Pampa alta	1	7	17.0	126.65	Ventura	12	5	Minima	330.56	Normal	132.23	-847.59	14.72	122%	Regular MO	NaN



```
df2=(df[(df.Sobrecosto=='Con SC') | (df.Sobrecosto=='SC elevado')])
df2
```

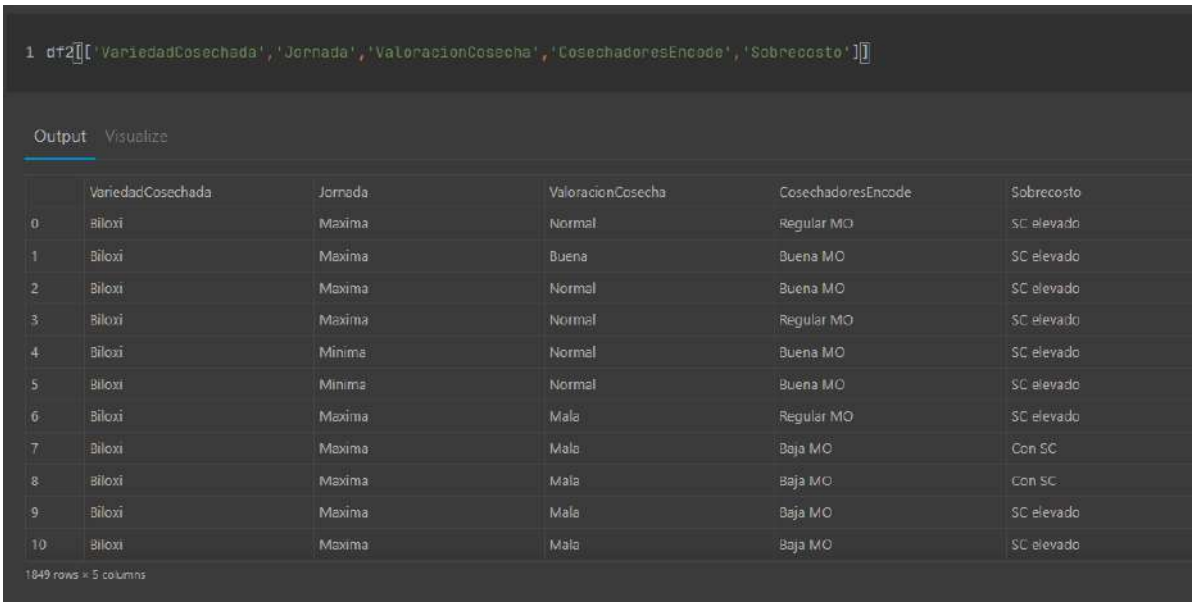
Output Visualize

	rasTrabajadas	Jornada	KilosCosechados	ValoracionCose...	CostoProducci...	DeficitProducci...	CostoKia (S/.)	Cumplimiento (...	CosechadoresE...	Sobrecosto
0		Maxima	510.05	Normal	2004.5	362.14	24.99	59%	Regular MO	SC elevado
1		Maxima	873.98	Buena	3434.73	620.52	42.82	59%	Buena MO	SC elevado
2		Maxima	719.58	Normal	2827.96	510.9	35.26	59%	Buena MO	SC elevado
3		Maxima	317.06	Normal	1246.04	225.11	15.54	59%	Regular MO	SC elevado
4		Minima	607.33	Normal	2386.8	431.2	29.76	59%	Buena MO	SC elevado
5		Minima	554.52	Normal	2179.25	393.71	27.17	59%	Buena MO	SC elevado
6		Maxima	215.59	Mala	847.28	153.07	10.56	59%	Regular MO	SC elevado
7		Maxima	45.46	Mala	178.64	32.27	2.23	59%	Baja MO	Con SC
8		Maxima	54.55	Mala	214.37	38.73	2.67	59%	Baja MO	Con SC
9		Maxima	145.46	Mala	571.66	103.28	7.13	59%	Baja MO	SC elevado
10		Maxima	166.24	Mala	653.33	118.03	8.15	59%	Baja MO	SC elevado

1849 rows × 21 columns

4. Exportar la data final con la que se trabajará y procesará en los softwares de Weka en su versión 3.8.5 y Knime en su versión 4.4.1.

Imagen 84: Exportación de la data final para árbol de decisión



```
df2[['VariedadCosechada','Jornada','ValoracionCosecha','CosechadoresEncode','Sobrecosto']]
```

Output Visualize

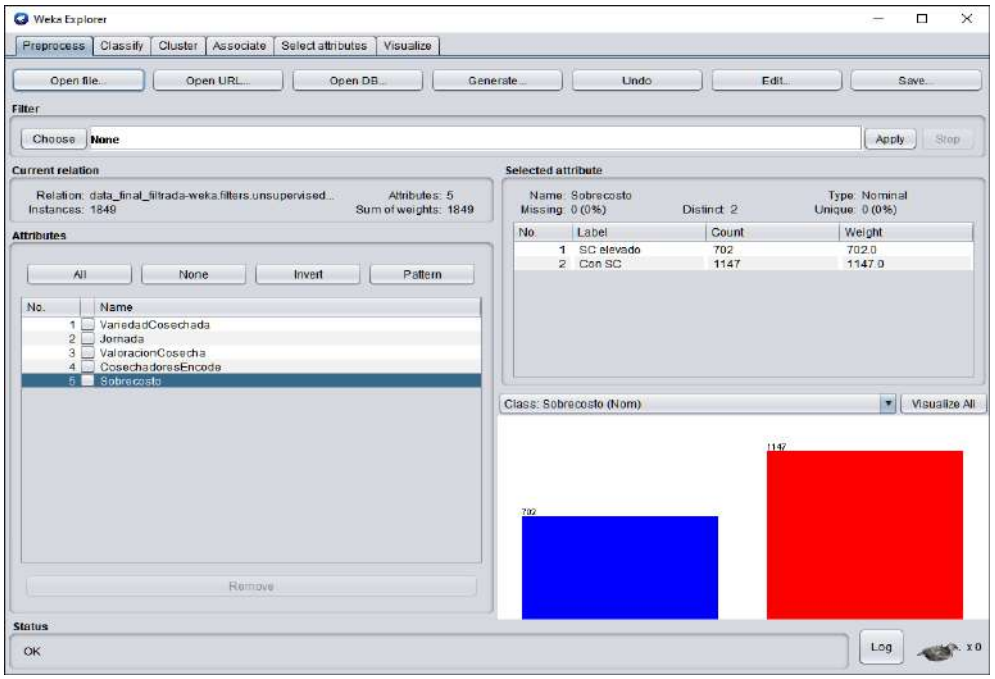
	VariedadCosechada	Jornada	ValoracionCosecha	CosechadoresEncode	Sobrecosto
0	Biloxi	Maxima	Normal	Regular MO	SC elevado
1	Biloxi	Maxima	Buena	Buena MO	SC elevado
2	Biloxi	Maxima	Normal	Buena MO	SC elevado
3	Biloxi	Maxima	Normal	Regular MO	SC elevado
4	Biloxi	Minima	Normal	Buena MO	SC elevado
5	Biloxi	Minima	Normal	Buena MO	SC elevado
6	Biloxi	Maxima	Mala	Regular MO	SC elevado
7	Biloxi	Maxima	Mala	Baja MO	Con SC
8	Biloxi	Maxima	Mala	Baja MO	Con SC
9	Biloxi	Maxima	Mala	Baja MO	SC elevado
10	Biloxi	Maxima	Mala	Baja MO	SC elevado

1849 rows × 5 columns

5. Implementamos el árbol de decisión en el software Weka en su versión 3.8.5, para lo cual también seguimos ciertos pasos que mostramos a continuación:

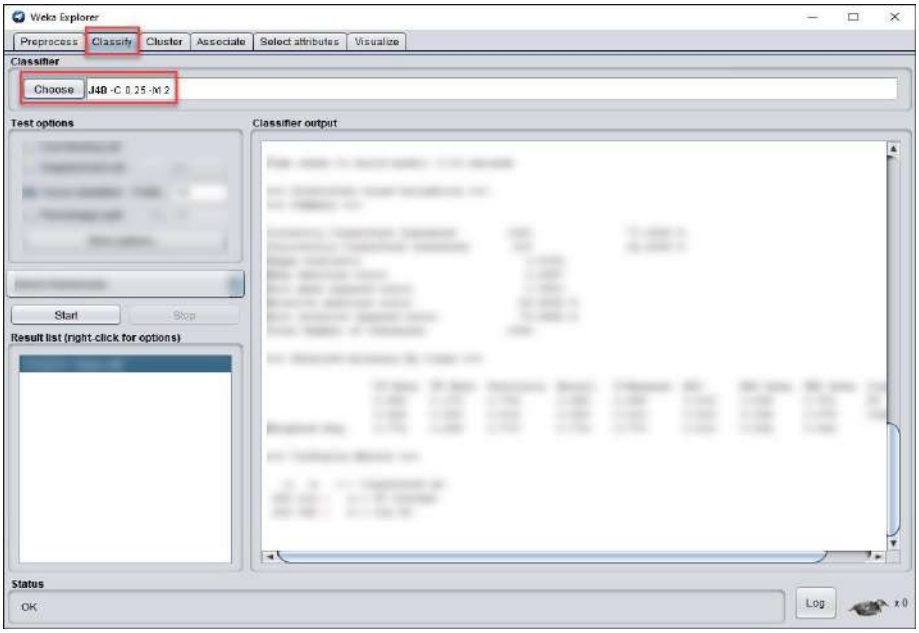
   5.1. Ingresamos los datos en archivo .csv al software Weka.

Imagen 85: Ingreso de datos al software Weka



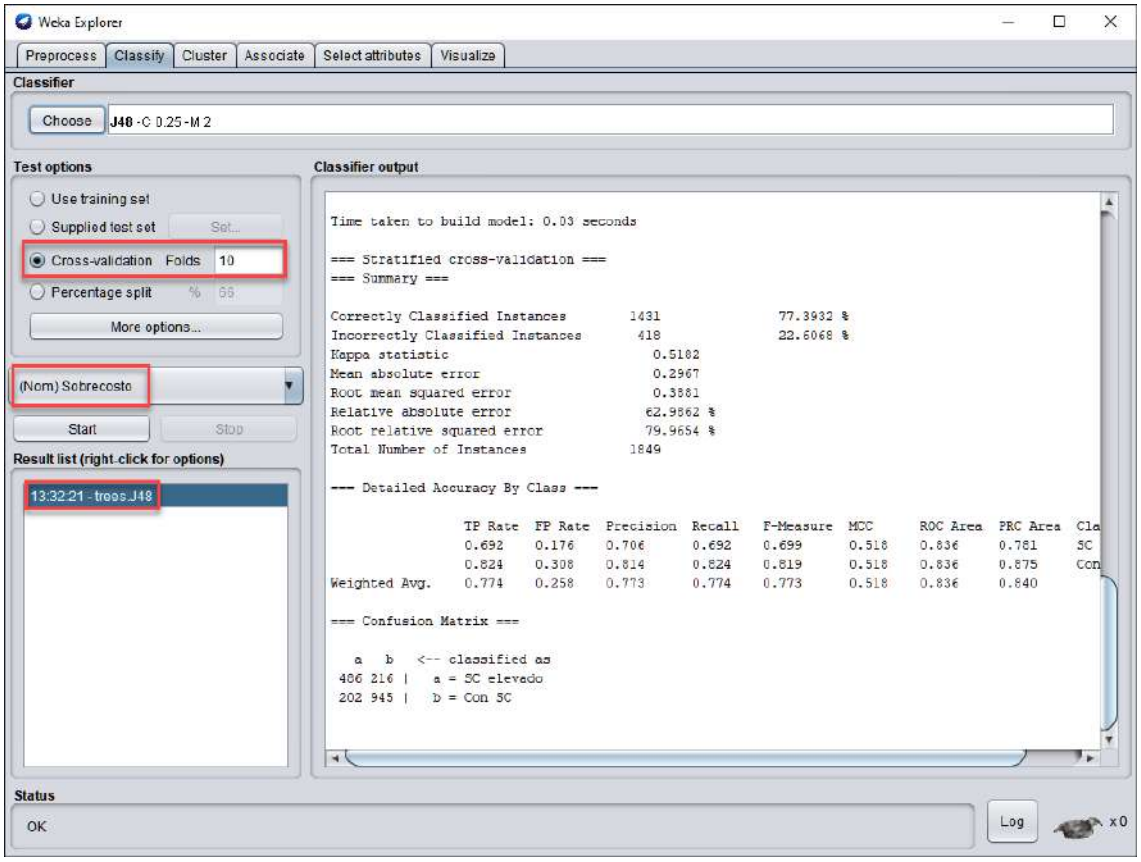
5.2. Seleccionamos la pestaña de “Classify”; para nuestro en particular seleccionar el “Choose” a “decisión tree J48”, mismo que implementa el algoritmo ID3.

Imagen 86: Elección del árbol de decisión en software Weka



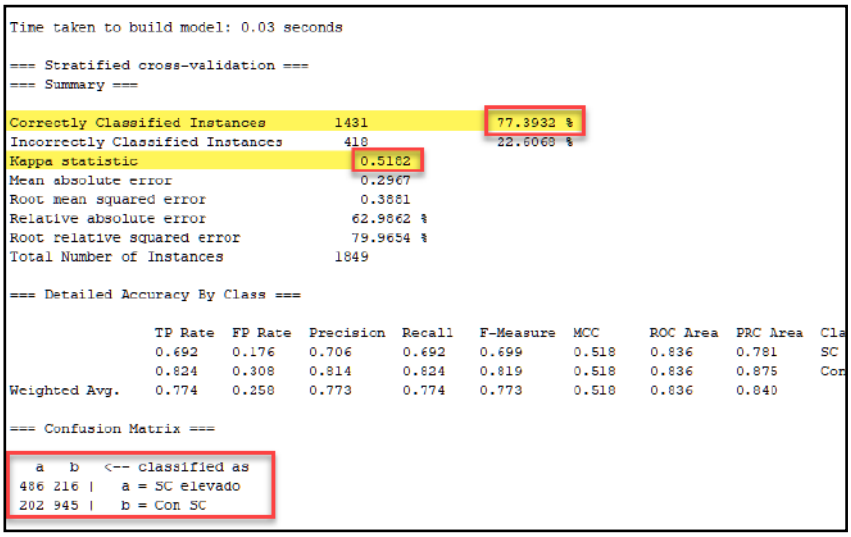
5.3. Ahora seleccionamos la validación cruzada y considera a 10 folds (grupos de muestras). Tras ello, el mismo software es quien considera la variable más apropiada para la clasificación, en este caso se coloca a “sobrecosto”; concluir con darle a “Start”.

Imagen 87: Configuración del árbol de decisión en Weka



5.4. El software Weka nos arroja estos resultados mostrados en la siguiente imagen.

Imagen 88: Resultados del árbol de decisión en Weka



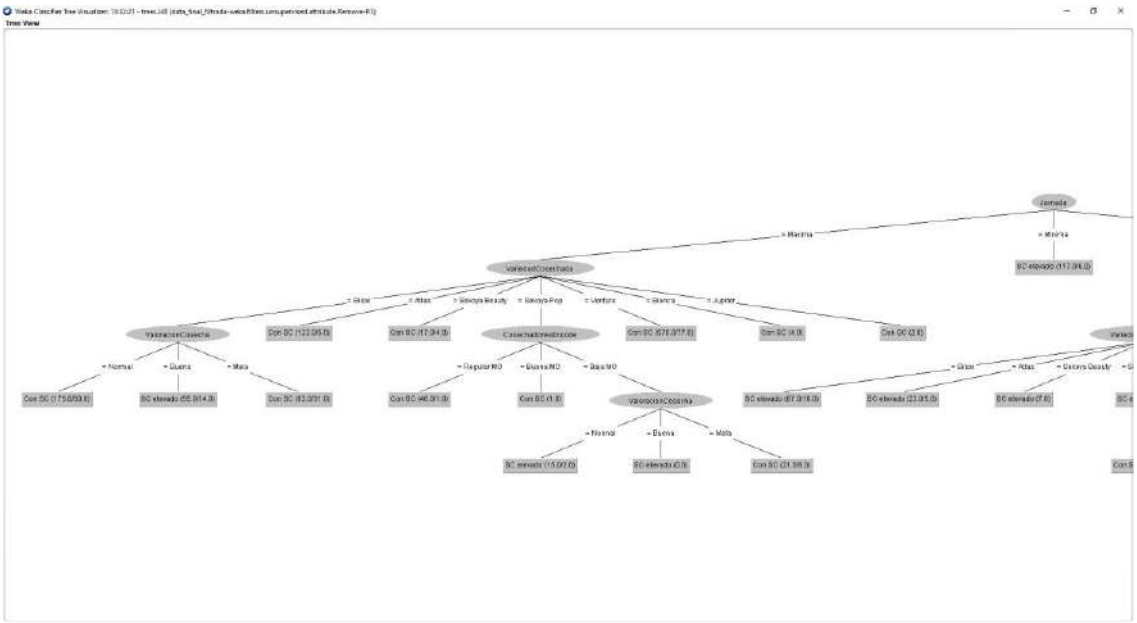
```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1431               77.3932 %
Incorrectly Classified Instances       418               22.6068 %
Kappa statistic                          0.5182
Mean absolute error                      0.2967
Root mean squared error                  0.3861
Relative absolute error                 62.9862 %
Root relative squared error             79.9654 %
Total Number of Instances             1849

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Cla
                 0.692    0.176      0.706     0.692     0.699     0.518    0.836     0.781     SC
                 0.824    0.308      0.814     0.824     0.819     0.518    0.836     0.875     Con
Weighted Avg.    0.774    0.258      0.773     0.774     0.773     0.518    0.836     0.840

=== Confusion Matrix ===

   a   b   <-- classified as
 486 216 |   a = SC elevado
 202 945 |   b = Con SC
```

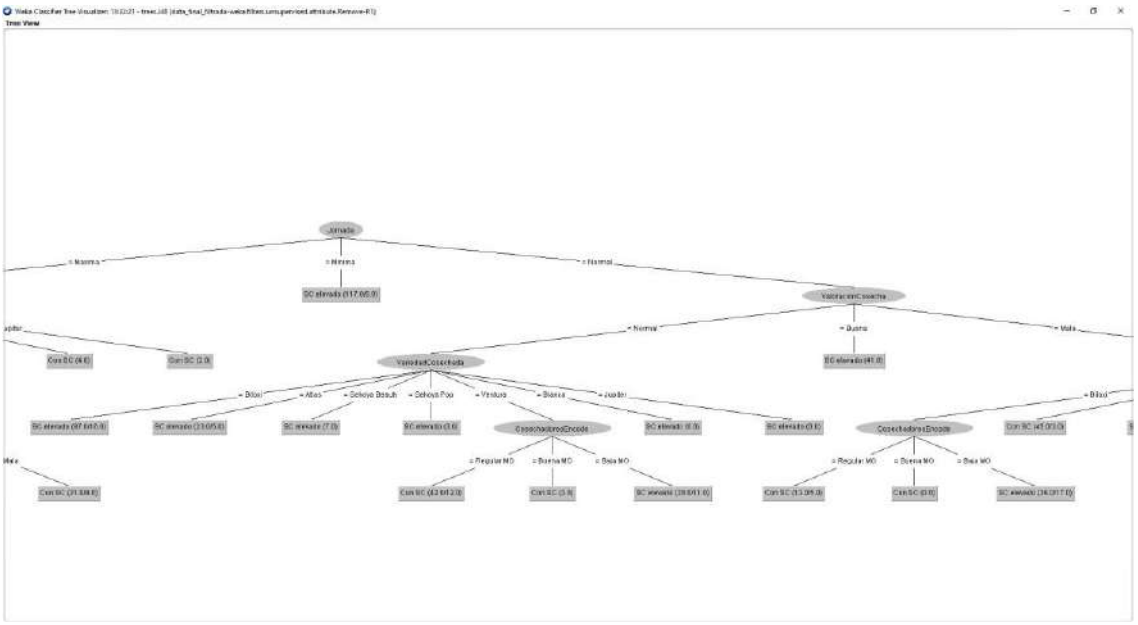
Los resultados se interpretan de la siguiente manera:

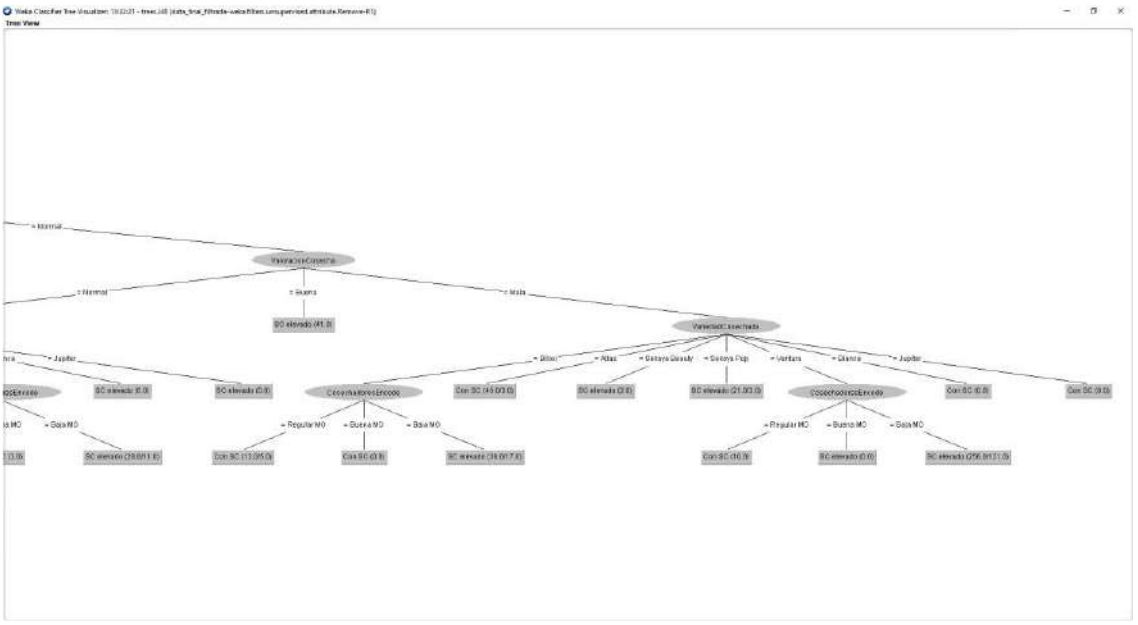
- El árbol de decisión tiene un “Accuracy” del 77,3932%.
- La coherencia entre las variables que consideramos para el árbol de decisión (Kappa statistic) es del 51,82%.
- Al final, se muestra la matriz de confusión.

5.5. Por último, mostramos el árbol de decisión como tal.

Imagen 89: Árbol de decisión en el software Weka.



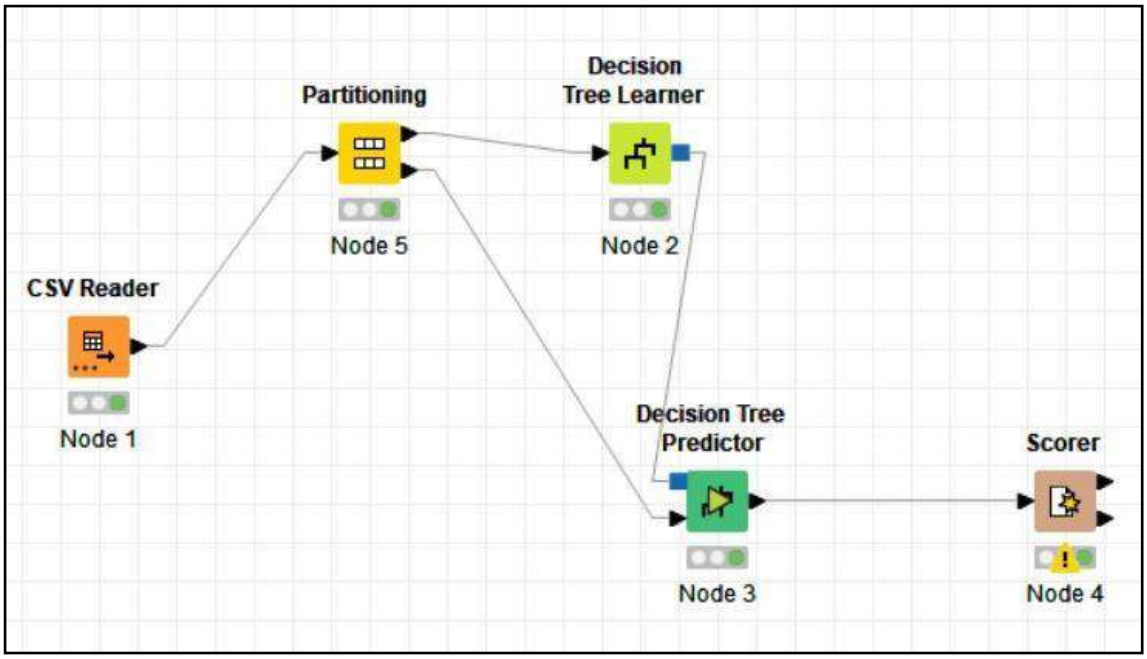




6. Implementamos, como último, el árbol de decisión en el software Knime en su versión 4.4.1, para lo cual también seguimos ciertos pasos que mostramos a continuación:

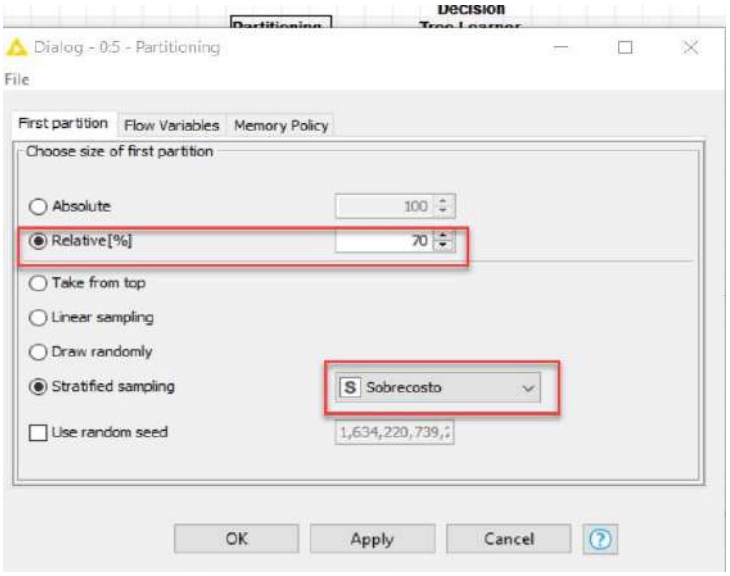
   6.1. De acuerdo con lo visto en el software Weka, y aprovechando las opciones que Knime ofrece, decidimos incluir un “Partitioning” dividendo así el conjunto de datos para entrenamiento y prueba; quedando el modelo como se muestra en la siguiente imagen.

Imagen 90: Modelo de árbol de decisión con "Partitioning" en Kinime



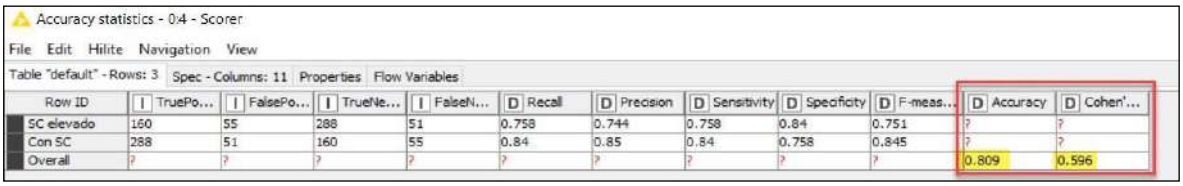
6.2. Colocamos en 70% como entrenamiento y dejamos un 30% como prueba; eligiendo clase predictora al sobrecosto.

Imagen 91: Configuración de árbol de decisión en el software Knime



6.3. Corroboramos los datos y mostramos los resultados que se ven en la siguiente imagen.

Imagen 92: Resultados del árbol de decisión en Knime



Accuracy statistics - 0:4 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
SC elevado	160	55	288	51	0.758	0.744	0.758	0.84	0.751	?	?
Con SC	288	51	160	55	0.84	0.85	0.84	0.758	0.845	?	?
Overall	?	?	?	?	?	?	?	?	?	0.809	0.596

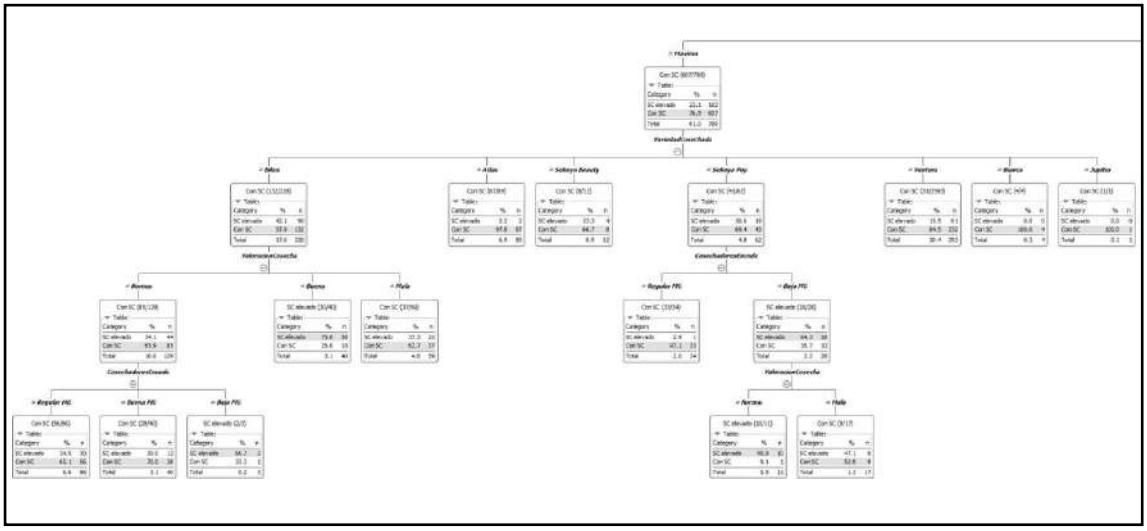
Los resultados son interpretados de la siguiente manera:

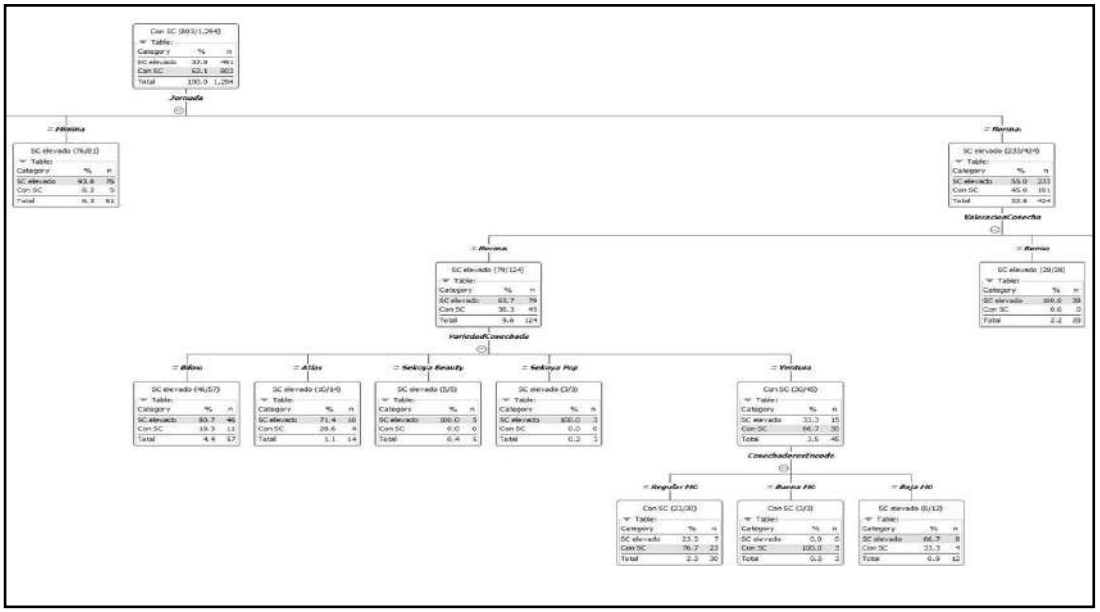
- Una exactitud o “accuracy” del árbol de decisión incrementó al 80,9% haciendo uso de la configuración y opciones que ofrece el software Knime.
- Una coherencia entre las variables que consideramos para el árbol de decisión, también se vio incrementada con respecto a la precisión hasta un 59,6%.

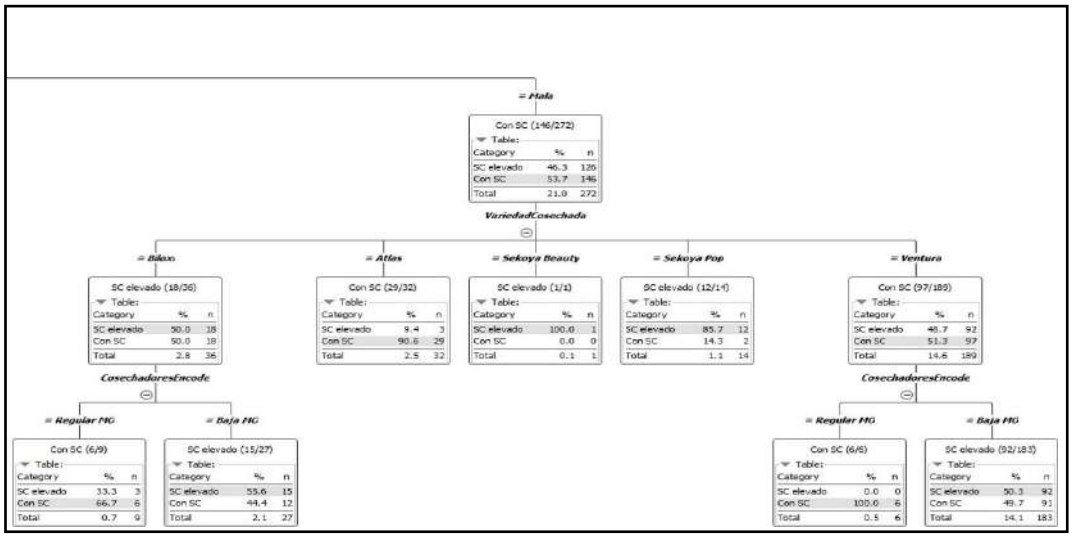
Por tanto, es favorable para la empresa y el trabajo de investigación **considerar como mejor árbol de decisión el obtenido en el software Knime en su versión 4.4.1.**

6.4. Finalmente, mostramos el árbol tal cual lo arroja Knime.

Imagen 93: Árbol de decisión en el software Knime.




Con SC (803/1,294)
Jornada
= Mañana
SC elevado (76/81)
= Tarde
SC elevado (233/424)
VariedadCosechada
= Bárex
SC elevado (46/57)
= Atlas
SC elevado (10/14)
= Sekoya Beauty
SC elevado (1/1)
= Sekoya Pop
SC elevado (3/3)
= Ventura
Con SC (30/40)
CosechadoresEncode
= Regular MG
Con SC (23/30)
= Buena MG
Con SC (3/3)
= Baja MG
SC elevado (6/12)


= Mala
Con SC (146/272)
Category % n
SC elevado 46.3 126
Con SC 53.7 146
Total 21.0 272
VariedadCosechada
= Bárex
SC elevado (18/36)
Con SC 50.0 18
SC elevado 50.0 18
Total 2.8 36
= Atlas
Con SC (29/32)
SC elevado 9.4 3
Con SC 90.6 29
Total 2.5 32
= Sekoya Beauty
SC elevado (1/1)
SC elevado 100.0 1
Con SC 0.0 0
Total 0.1 1
= Sekoya Pop
SC elevado (12/14)
SC elevado 85.7 12
Con SC 14.3 2
Total 1.1 14
= Ventura
Con SC (97/189)
SC elevado 48.7 92
Con SC 51.3 97
Total 14.6 189
CosechadoresEncode
= Regular MG
Con SC (6/9)
SC elevado 33.3 3
Con SC 66.7 6
Total 0.7 9
= Baja MG
SC elevado (15/27)
SC elevado 55.6 15
Con SC 44.4 12
Total 2.1 27
= Regular MG
Con SC (6/6)
SC elevado 0.0 0
Con SC 100.0 6
Total 0.5 6
= Baja MG
SC elevado (92/183)
SC elevado 50.3 92
Con SC 49.7 91
Total 14.1 183

7. Obtuvimos las estadísticas y matriz de confusión.

Tabla 21: Estadísticas de precisión árbol de decisión por Knime

	True Positives	***False Positives***	***True Negatives***	***False Negatives***	***Recall***	***Precision***	***Sensitivity***	***Specificity***	***F-Measure***	***Accuracy***	***Cohen's kappa***
SC elevado	160	55	288	51	0.758	0.744	0.758	0.84	0.751	?	?
Con SC	288	51	160	55	0.84	0.85	0.84	0.758	0.845	?	?
Overall	?	?	?	?	?	?	?	?	?	0.809	0.596

Tabla 22: Matriz de confusión árbol de decisión por Knime

	SC elevado	***Con SC***
SC elevado	160	51
Con SC	55	288

Los resultados del árbol de decisión nos muestran algunos resultados como los siguientes:

- Si la jornada laboral es máxima con una cosecha de arándano de tipo Biloxi y una valoración de cosecha buena se obtiene que existe un 75% de probabilidades de que se obtenga un sobrecosto elevado (SC elevado).
- Si la jornada laboral es máxima con una cosecha de arándano de tipo Ventura, sin otras características adicionales, obtenemos que existe un 15.5% de probabilidades de que se obtenga un sobrecosto elevado (SC elevado).
- Si la jornada laboral es mínima, sin otras características adicionales, se obtiene que existe un 93.6% de probabilidades de que habrá un sobrecosto elevado (SC elevado).

Así como los casos anteriores, tenemos algunos otros que se obtiene una seguridad al 100% de predecir que se obtendrá un sobrecosto elevado (SC elevado), mismos que son los siguientes:

- **Caso 1:** Se predice en un 100% que, si la jornada laboral es normal y la valoración de la cosecha es buena, entonces se tendrá un sobrecosto elevado. Una **solución propuesta** por nosotros sería observar con detalle los obreros empleados en estas jornadas normales y los horarios en detalle para tomar mejores decisiones.

- **Caso 2:** Se predice en un 100% que, si la jornada laboral es normal y la valoración de la cosecha es normal, las variedades de arándanos cosechadas como Biloxi, Atlas, Sekoya Beauty, Sekoya Pop tendrán sobrecosto elevado, exceptuando a Ventura que obtiene un 66.7% de probabilidades de solo tener sobrecosto. Una **solución propuesta** por nosotros sería el detalle de los recursos colocados en estas variedades y qué diferencia se observa de esto.

4.2.3. RESULTADO 3 vs OE3: IMPLEMENTAR BINNING Y REGLAS DE ASOCIACIÓN CON FIN DE HALLAR LAS RELACIONES ENTRE VALORES AGRUPADOS CON RESPECTO AL CUMPLIMIENTO DE KIAS.

Para realizar la implementación del árbol de decisión se siguieron una serie de pasos mostrados a continuación:

1. Identificamos las variables que usaremos:
    - Cumplimiento (%).
    - Kilos cosechados.
    - Variedad cosechada.
    - Costo de Kia.

2. Tras haber identificado las columnas a usar de la data consolidada original, entonces es momento de pasarlas a columnas nominales, quedarían de la siguiente forma:

    - **Variedad cosechada:** Esta columna, desde la data consolidada original, se compone por clases lo que la hace nominal desde el inicio y se categoriza de la siguiente manera:
        - Biloxi.
        - Atlas.
        - Sekoya Beauty.
        - Sekoya Pop.
        - Ventura.
        - Bianca.
        - Jupiter.
    - **Kilos cosechados:** Haremos uso de la categorización que realizamos para el OE1 y quedaba de la siguiente manera:
        - **Mala:** si los kilos cosechados fueron <=250.
        - **Normal:** si los kilos cosechados fueron >250 y <800.
        - **Bueno:** si los kilos cosechados fueron >800.
    - **Cumplimiento (%):** Esta columna indicaba el porcentaje con respecto al cumplimiento de Kias durante un periodo determinado de tiempo, que este caso se medía por semanas. Haciendo uso del Binning que se explicará en detalle más adelante, la categorización quedaría de la siguiente manera:
        - **Sin cumplir:** si el cumplimiento fue de <95%.

- **Incompleto:** si el cumplimiento fue >=95% y <=100%.
- **Cumplida:** si el cumplimiento fue >100%.

- **CostoKia (S/.):** Esta columna indica el costo de la Kia por cosecha. Nuevamente haciendo uso del Binning, la columna nominal queda categorizada de la siguiente manera:
  - **Costo PPTO (costo presupuestado):** si el costo fue <30 soles.
  - **Costo NoPPTO (costo no presupuestado):** si el costo fue >=30 y <=50 soles.
  - **Costo Elevado:** si el costo fue >50 soles.

3. La transformación a variables nominales la realizamos nuevamente en el lenguaje Python (preprocesamiento); sin embargo, específicamente para este objetivo realizamos la técnica del Binning, que se rige por usar puntos de cortes y el máximo/mínimo de la columna numérica. Esto se hizo haciendo uso solo de la librería pandas, tal cual se muestra en las imágenes.

Imagen 94: Creación de función Binning


```
[ ]  1 def funcionBinning(col, puntos_corte, labels=None):
     2     nivel_min=col.min()
     3     nivel_max=col.max()
     4     break_points=[nivel_min] + puntos_corte + [nivel_max]
     5     print(break_points)
     6
     7     if not labels:
     8         labels=range(len(puntos_corte)+1)
     9     colBin=pd.cut(col, bins=break_points, labels=labels, include_lowest=True)
    10     return colBin
```

4. A continuación, mostraremos el preprocesamiento de las variables numéricas a nominales.

4.1. Para Cumplimiento (%):

Imagen 95: Preprocesamiento columna Cumplimiento (%)

```
[ ]  1 puntos_corte = [95,100]
     2 labels = ["Sin cumplir","Incompleto","Cumplido"]
     3 df2["cumplimientoEncode"] = funcionBinning(df2["Cumplimiento ()"], puntos_corte, labels)
```

	Unnamed: 0	Semana	VariedadCosechada	KilosCosechados	CostoKia (S/.)	Cumplimiento ()	ValoracionCosecha	cumplimientoEncode
0	0	29	Biloxi	510.05	24.99	59	Normal	Sin cumplir
1	1	29	Biloxi	873.98	42.82	59	Buena	Sin cumplir
2	2	29	Biloxi	719.58	35.26	59	Normal	Sin cumplir
3	3	29	Biloxi	317.06	15.54	59	Normal	Sin cumplir
4	4	29	Biloxi	607.33	29.76	59	Normal	Sin cumplir
...	...	...	...	...	...	...	...	...
17962	17962	49	Biloxi	144.52	7.08	122	Mala	Cumplido
17963	17963	49	Ventura	262.76	12.88	122	Normal	Cumplido
17964	17964	49	Ventura	373.39	18.30	122	Normal	Cumplido
17965	17965	49	Ventura	399.17	19.56	122	Normal	Cumplido
17966	17966	49	Ventura	300.56	14.73	122	Normal	Cumplido

17967 rows × 8 columns

4.2. Para CostoKia (S/.):

Imagen 96: Preprocesamiento para columna CostoKia(S/.)

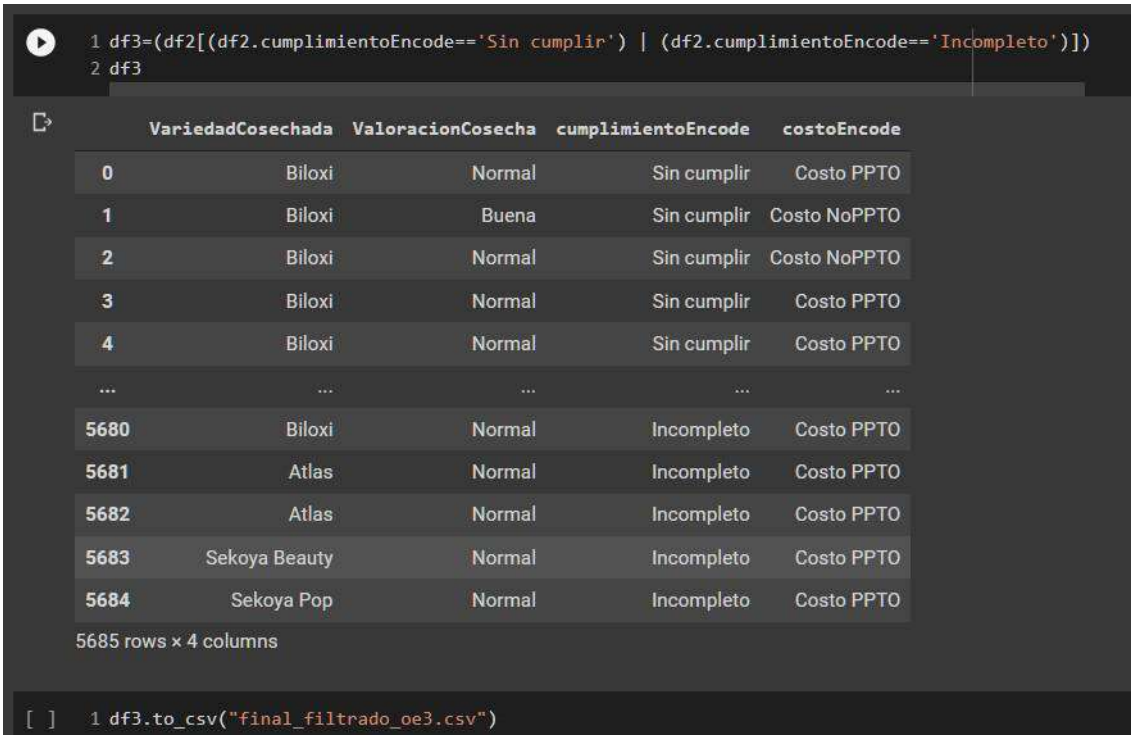
```
[ ]  1 puntos_corte = [30,50]
     2 labels = ["Costo PPTO","Costo NoPPTO","Costo Elevado"]
     3 df2["costoEncode"] = funcionBinning(df2["CostoKia (S/.)"], puntos_corte, labels)

     [0.0, 30, 50, 180.25]
```

	Unnamed: 0	VariedadCosechada	KilosCosechados	CostoKia (S/.)	Cumplimiento ()	ValoracionCosecha	cumplimientoEncode	costoEncode
0	0	Biloxi	510.05	24.99	59	Normal	Sin cumplir	Costo PPTO
1	1	Biloxi	873.98	42.82	59	Buena	Sin cumplir	Costo NoPPTO
2	2	Biloxi	719.58	35.26	59	Normal	Sin cumplir	Costo NoPPTO
3	3	Biloxi	317.06	15.54	59	Normal	Sin cumplir	Costo PPTO
4	4	Biloxi	607.33	29.76	59	Normal	Sin cumplir	Costo PPTO
...	...	...	...	...	...	...	...	...
17962	17962	Biloxi	144.52	7.08	122	Mala	Cumplido	Costo PPTO
17963	17963	Ventura	262.76	12.88	122	Normal	Cumplido	Costo PPTO
17964	17964	Ventura	373.39	18.30	122	Normal	Cumplido	Costo PPTO
17965	17965	Ventura	399.17	19.56	122	Normal	Cumplido	Costo PPTO
17966	17966	Ventura	300.56	14.73	122	Normal	Cumplido	Costo PPTO

17967 rows × 8 columns

5. Sabiendo que las demás variables ya se tienen resueltas en categorizadas (nominales), solo queda por acotar que de aquella categorización que se realizó para el cumplimiento, ´se tomará en cuenta nada más aquellas variables con etiquetas “Sin cumplir” e “Incompleto”, ya que, nos enfocamos en el problema de fondo.

Imagen 97: Filtrado de data final



```
1 df3=(df2[(df2.cumplimientoEncode=='Sin cumplir') | (df2.cumplimientoEncode=='Incompleto')])
2 df3
```

	VariedadCosechada	ValoracionCosecha	cumplimientoEncode	costoEncode
0	Biloxi	Normal	Sin cumplir	Costo PPTO
1	Biloxi	Buena	Sin cumplir	Costo NoPPTO
2	Biloxi	Normal	Sin cumplir	Costo NoPPTO
3	Biloxi	Normal	Sin cumplir	Costo PPTO
4	Biloxi	Normal	Sin cumplir	Costo PPTO
...	...	...	...	...
5680	Biloxi	Normal	Incompleto	Costo PPTO
5681	Atlas	Normal	Incompleto	Costo PPTO
5682	Atlas	Normal	Incompleto	Costo PPTO
5683	Sekoya Beauty	Normal	Incompleto	Costo PPTO
5684	Sekoya Pop	Normal	Incompleto	Costo PPTO

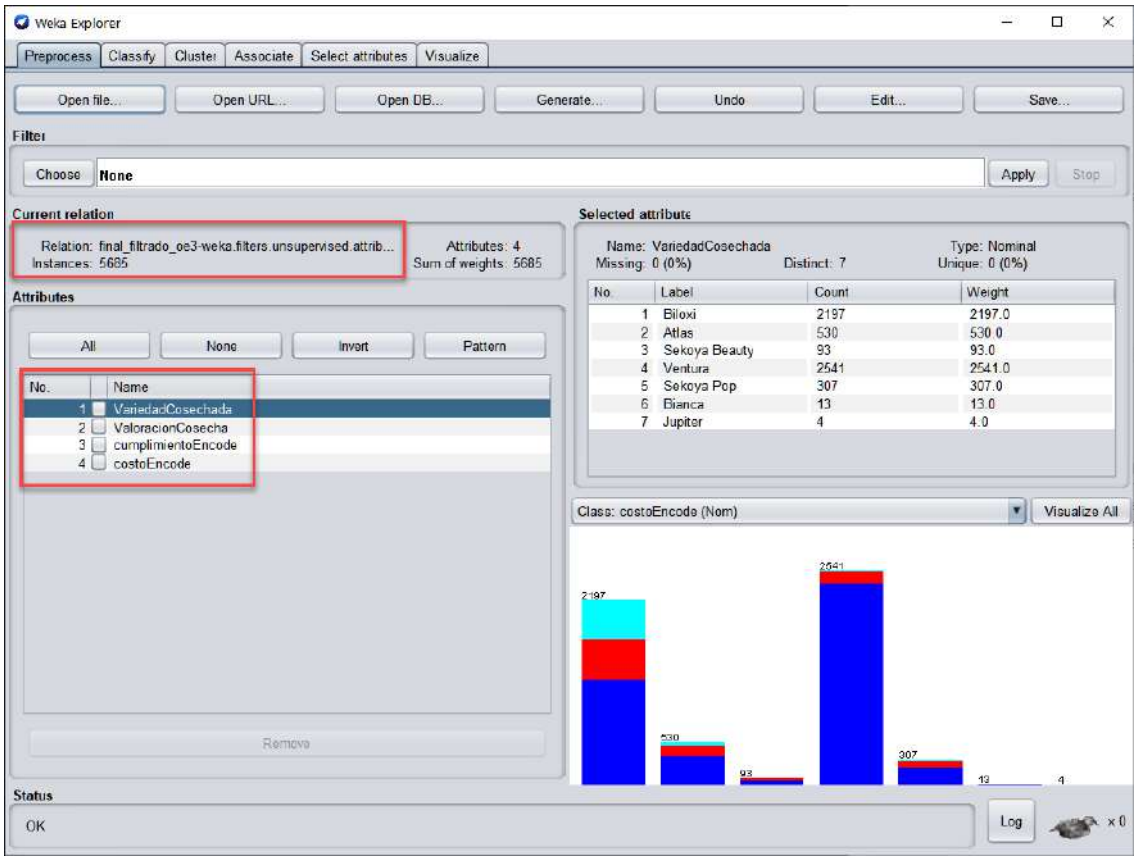
5685 rows × 4 columns

```
[ ]  1 df3.to_csv("final_filtrado_oe3.csv")
```

6. En principio, usamos Weka en su versión 3.8.5, haciendo uso de la técnica A priori para hallar las reglas de asociación.

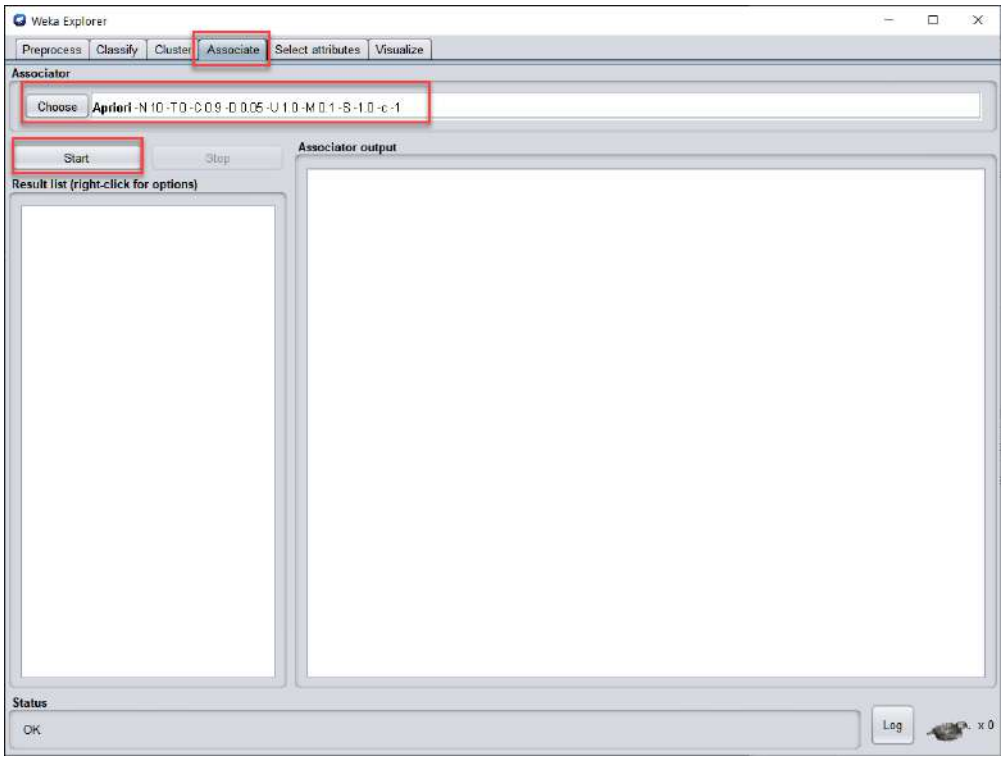
   I. Paso 1: Ingresamos al software, elegimos él .csv con solo las variables que consideramos en el preprocesamiento.

Imagen 98: Configuración inicial del software Weka para OE3



II. Paso 2: Seleccionamos la pestaña de “Associate” y elegimos la técnica “A priori” y damos en el botón Start.

Imagen 99: Configuración técnica de asociación Weka para OE3



III. Paso 3: Destaca una serie de variables una vez dimos a Start, de las cuales podemos resaltar las reglas encontradas, así como otras características.

Imagen 100: Resultados de las reglas de asociación a priori Weka para OE3

IV. Paso 4: Se destacan las siguientes características:

Imagen 101: Resultados de las reglas de asociación a priori Weka OE3

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     final_filtrado_oe3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    5685
Attributes:   4
              VariedadCosechada
              ValoracionCosecha
              cumplimientoEncode
              costoEncode
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (569 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 12

Size of set of large itemsets L(4): 1

Best rules found:

1. ValoracionCosecha=Mala 2263 ==> costoEncode=Costo PPTO 2263    <conf:(1)> lift:(1.33) lev:(0.1) [561] conv:(561.27)
2. ValoracionCosecha=Mala cumplimientoEncode=Sin cumplir 1679 ==> costoEncode=Costo PPTO 1679    <conf:(1)> lift:(1.33) lev:(0.07) [416] conv:(416.43)
3. VariedadCosechada=Ventura ValoracionCosecha=Mala 1494 ==> costoEncode=Costo PPTO 1494    <conf:(1)> lift:(1.33) lev:(0.07) [370] conv:(370.56)
4. VariedadCosechada=Ventura ValoracionCosecha=Mala cumplimientoEncode=Sin cumplir 1109 ==> costoEncode=Costo PPTO 1109    <conf:(1)> lift:(1.33) lev:(0.05) [275] conv:(275.06)
5. ValoracionCosecha=Mala cumplimientoEncode=Incompleto 584 ==> costoEncode=Costo PPTO 584    <conf:(1)> lift:(1.33) lev:(0.03) [144] conv:(144.84)
6. VariedadCosechada=Ventura cumplimientoEncode=Sin cumplir 1568 ==> costoEncode=Costo PPTO 1543    <conf:(0.97)> lift:(1.29) lev:(0.06) [368] conv:(8.56)
7. VariedadCosechada=Ventura 2541 ==> costoEncode=Costo PPTO 2383    <conf:(0.94)> lift:(1.25) lev:(0.08) [472] conv:(3.96)
```

Entre ellas podemos destacar siguiente:

- El soporte mínimo es de 0.1 (1% en otros softwares).
- La confidencialidad (métrica considerada) mínima es de 0.9, lo cual indica que las reglas tienen una exactitud considerablemente alta.

V. Paso 5: A continuación, en las siguientes tablas se presentan las reglas arrojadas por el software Weka:

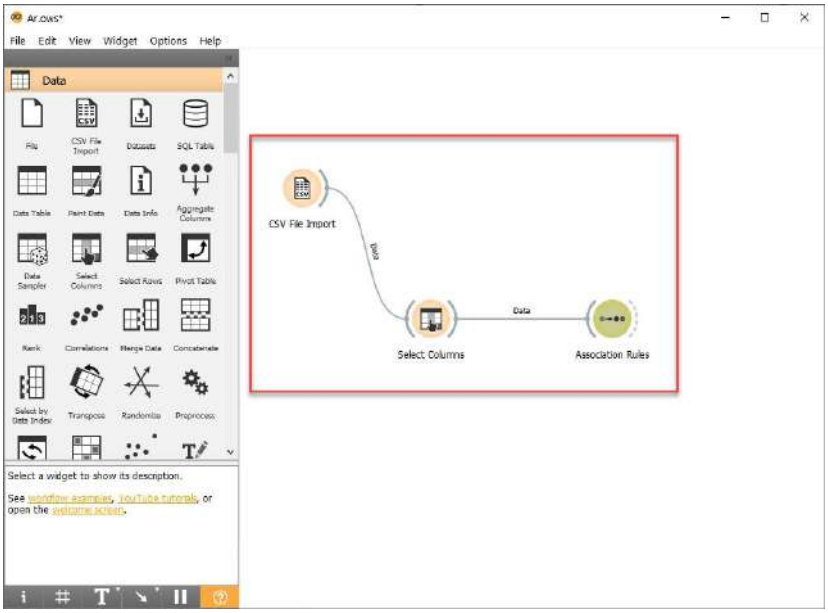
Tabla 23: Reglas de asociación con técnica A priori (Weka)

	Antecedente	***Consecuente***	***Conf***	***Lift***
1	ValoracionCosecha=Mala cumplimientoEncode=Sin Cumplir	CostoEncode=Costo PPTO	1.00	1.33
2	VariedadCosechada=Ventura ValoraciónCosecha=Mala CumplimientoEncode=Sin cumplir	CostoEncode=Costo PPTO	1.00	1.33
3	VariedadCosechada=Ventura cumplimientoEncode=Incompleto	CostoEncode=Costo PPTO	1.00	1.33
4	VariedadCosechada=Ventura cumplimientoEncode=Sin Cumplir	CostoEncode=Costo PPTO	0.97	1.29

7. Tras observar que Weka no arrojaba con la técnica A priori, la cantidad de reglas requeridas; optamos por otro software, en este caso Orange en su versión 3.30.1, mismo trabaja con la técnica FP-Growth para las reglas de asociación.

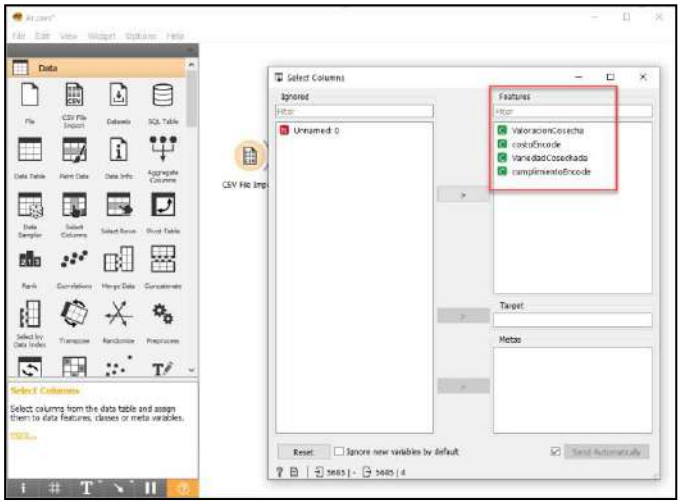
I. Paso 1: Abrimos el software Orange y realizamos el modelo de para obtener las reglas de asociación.

Imagen 102: Creación modelo regla de asociación Orange para OE3



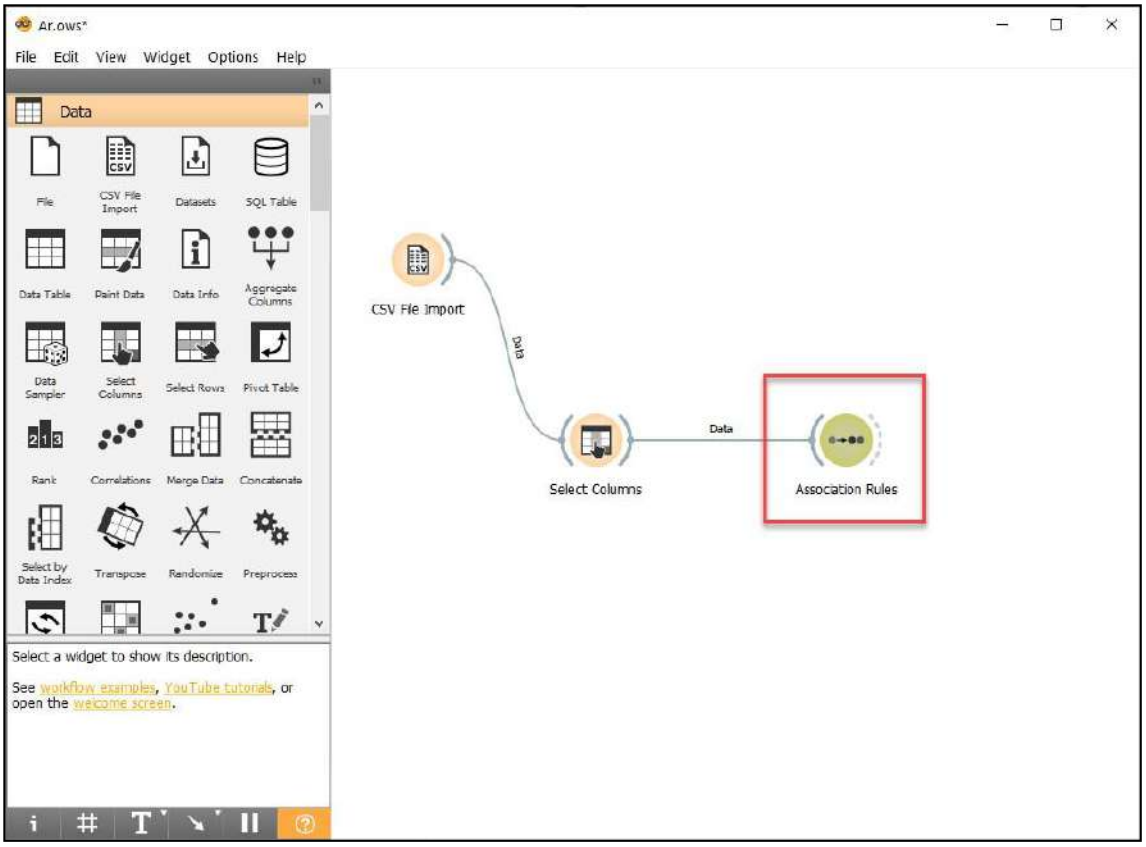
II. Paso 2: Seleccionamos las columnas que usaremos, en este caso solo descartamos la columna “Unnamed”.

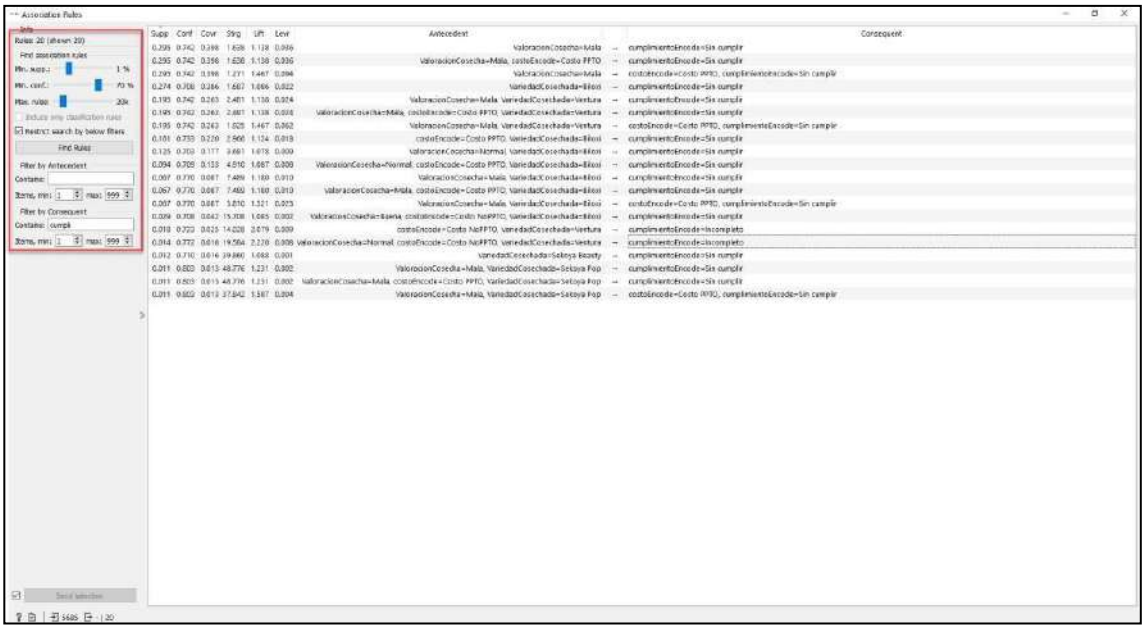
Imagen 103: Selección de los datos para OE3



III. Paso 3: Ingresamos “Association Rules” y configuramos según sea nuestros requerimientos; en nuestro caso consideramos hasta un 70% de confidencialidad, el soporte lo dejamos en 1%.

Imagen 104: Ejecución del modelo propuesto en Orange para OE3





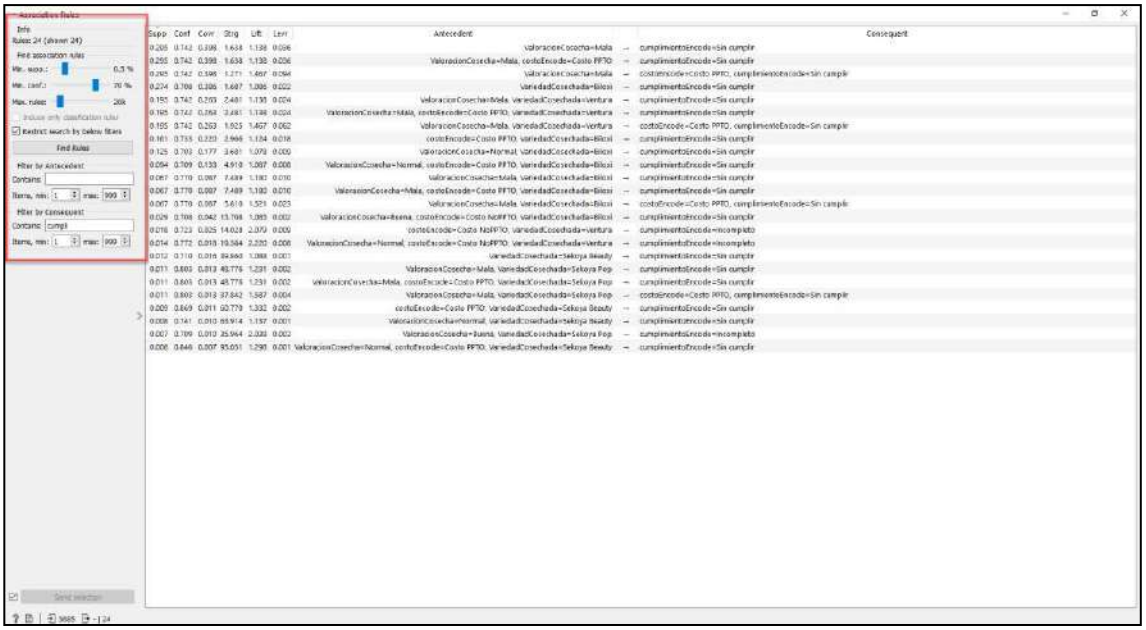
Nos arroja las siguientes reglas:

Imagen 105: Primeros resultados reglas de asociación Orange OE3

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.011	0.803	0.013	48.776	1.231	0.002	ValoracionCosecha=Mala, VariedadCosechada=Sekoya Pop	→	cumplimientoEncode=Sin cumplir
0.011	0.803	0.013	48.776	1.231	0.002	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Sekoya Pop	→	cumplimientoEncode=Sin cumplir
0.011	0.803	0.013	37.842	1.587	0.004	ValoracionCosecha=Mala, VariedadCosechada=Sekoya Pop	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.014	0.772	0.018	19.584	2.220	0.008	ValoracionCosecha=Normal, costoEncode=Costo NoPPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Incompleto
0.067	0.770	0.087	7.489	1.180	0.010	ValoracionCosecha=Mala, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.067	0.770	0.087	7.489	1.180	0.010	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.067	0.770	0.087	5.810	1.521	0.023	ValoracionCosecha=Mala, VariedadCosechada=Biloxi	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	2.481	1.138	0.024	ValoracionCosecha=Mala, VariedadCosechada=Ventura	→	cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	2.481	1.138	0.024	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	1.925	1.467	0.062	ValoracionCosecha=Mala, VariedadCosechada=Ventura	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.638	1.138	0.036	ValoracionCosecha=Mala	→	cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.638	1.138	0.036	ValoracionCosecha=Mala, costoEncode=Costo PPTO	→	cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.271	1.467	0.094	ValoracionCosecha=Mala	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.161	0.733	0.220	2.966	1.124	0.018	costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.018	0.723	0.025	14.028	2.079	0.009	costoEncode=Costo NoPPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Incompleto
0.012	0.710	0.016	39.860	1.088	0.001	VariedadCosechada=Sekoya Beauty	→	cumplimientoEncode=Sin cumplir
0.094	0.709	0.133	4.910	1.087	0.008	ValoracionCosecha=Normal, costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.274	0.708	0.386	1.687	1.086	0.022	VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.029	0.708	0.042	15.708	1.085	0.002	ValoracionCosecha=Buena, costoEncode=Costo NoPPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.125	0.703	0.177	3.681	1.078	0.009	ValoracionCosecha=Normal, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir

IV. Paso 4: Modificamos las configuraciones y ahora el soporte lo bajos a 0.5%, dejando la confidencialidad en 70%.

Imagen 106: Configuración de soporte y confidencialidad Orange OE3



Podemos observar que nos arroja las siguientes reglas y arroja unas cuantas más:

Imagen 107: Resultados finales reglas de asociación Orange OE3

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.009	0.869	0.011	60.770	1.332	0.002	costoEncode=Costo PPTO, VariedadCosechada=Sekoya Beauty	→	cumplimientoEncode=Sin cumplir
0.006	0.846	0.007	55.051	1.298	0.001	ValoracionCosecha=Normal, costoEncode=Costo PPTO, VariedadCosechada=Sekoya Beauty	→	cumplimientoEncode=Sin cumplir
0.011	0.803	0.013	48.776	1.231	0.002	ValoracionCosecha=Mala, VariedadCosechada=Sekoya Pop	→	cumplimientoEncode=Sin cumplir
0.011	0.803	0.013	48.776	1.231	0.002	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Sekoya Pop	→	cumplimientoEncode=Sin cumplir
0.011	0.803	0.013	37.842	1.567	0.004	ValoracionCosecha=Mala, VariedadCosechada=Sekoya Pop	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.014	0.772	0.018	19.584	2.220	0.008	ValoracionCosecha=Normal, costoEncode=Costo NoPPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Incompleto
0.067	0.770	0.087	7.489	1.180	0.010	ValoracionCosecha=Mala, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.067	0.770	0.087	7.489	1.180	0.010	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.067	0.770	0.087	5.810	1.521	0.023	ValoracionCosecha=Mala, VariedadCosechada=Biloxi	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	2.481	1.138	0.024	ValoracionCosecha=Mala, VariedadCosechada=Ventura	→	cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	2.481	1.138	0.024	ValoracionCosecha=Mala, costoEncode=Costo PPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Sin cumplir
0.195	0.742	0.263	1.925	1.467	0.062	ValoracionCosecha=Mala, VariedadCosechada=Ventura	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.636	1.138	0.036	ValoracionCosecha=Mala	→	cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.636	1.138	0.036	ValoracionCosecha=Mala, costoEncode=Costo PPTO	→	cumplimientoEncode=Sin cumplir
0.295	0.742	0.398	1.271	1.467	0.094	ValoracionCosecha=Mala	→	costoEncode=Costo PPTO, cumplimientoEncode=Sin cumplir
0.008	0.741	0.010	63.914	1.137	0.001	ValoracionCosecha=Normal, VariedadCosechada=Sekoya Beauty	→	cumplimientoEncode=Sin cumplir
0.161	0.733	0.220	2.966	1.124	0.018	costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.018	0.723	0.025	14.028	2.079	0.009	costoEncode=Costo NoPPTO, VariedadCosechada=Ventura	→	cumplimientoEncode=Incompleto
0.012	0.710	0.016	39.860	1.088	0.001	VariedadCosechada=Sekoya Beauty	→	cumplimientoEncode=Sin cumplir
0.007	0.709	0.010	35.964	2.038	0.003	ValoracionCosecha=Buena, VariedadCosechada=Sekoya Pop	→	cumplimientoEncode=Incompleto
0.094	0.709	0.133	4.910	1.087	0.008	ValoracionCosecha=Normal, costoEncode=Costo PPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.274	0.708	0.386	1.687	1.086	0.022	VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.029	0.708	0.042	15.708	1.085	0.002	ValoracionCosecha=Buena, costoEncode=Costo NoPPTO, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir
0.125	0.703	0.177	3.681	1.078	0.009	ValoracionCosecha=Normal, VariedadCosechada=Biloxi	→	cumplimientoEncode=Sin cumplir

V. Paso 5: A continuación, en las siguientes tablas se presentan las reglas arrojadas por el software Orange.

Tabla 24: Reglas de asociación con técnica FP-Growth (Orange)

	Antecedente	***Consecuente***	***Supp***	***Conf***	***Lift***
1	ValoracionCosecha=Normal CostoEncode=Costo PPTO VariedadCosechada= SekoyaBeauty	cumplimientoEncode=Sin cumplir	0.006	0.846	1.298
2	ValoracionCosecha=Mala CostoEncode=Costo PPTO VariedadCosechada=Sekoya Pop	cumplimientoEncode=Sin cumplir	0.011	0.803	1.231
3	ValoracionCosecha=Normal CostoEncode=Costo NoPPTO VariedadCosechada=Ventura	cumplimientoEncode=Incompleto	0.014	0.772	2.220
4	ValoracionCosecha=Mala CostoEncode=Costo PPTO VariedadCosechada=Biloxi	cumplimientoEncode=Sin cumplir	0.067	0.770	1.180
5	ValoracionCosecha=Mala CostoEncode=Costo PPTO VariedadCosechada=Ventura	cumplimientoEncode=Sin cumplir	0.195	0.742	1.138
6	CostoEncode=Costo PPTO VariedadCosechada=Biloxi	cumplimientoEncode=Sin cumplir	0.161	0.733	1.124
7	CostoEncode=Costo NoPPTO VariedadCosechada=Ventura	cumplimientoEncode=Incompleto	0.018	0.723	2.079

Con respecto a la interpretación de los resultados del cuadro anteriormente presentando, colocamos dos ejemplos puntuales para observar los patrones encontrados:

- **Caso 1:** Con un 84,6% de confianza se obtiene que se consigue un estado sin cumplir cuando la valoración de la cosecha es normal, el costo es presupuestado y la variedad cosechada es Sekoya Beauty.
- **Caso 2:** Con un 77,2% de confianza se obtiene que se consigue un estado incompleto cuando la valoración de la cosecha es normal, el costo no está presupuestado y la variedad cosechada es Ventura.

Del mismo modo existen otras reglas de asociación como los casos presentados, que detallan el problema del cumplimiento de Kias en ciertos escenarios con respecto a características específicas de la cosecha; sin embargo, la solución a estos y otros casos dependerá de la gestión interna que se realice con los datos arrojados por la técnica predictiva usada.

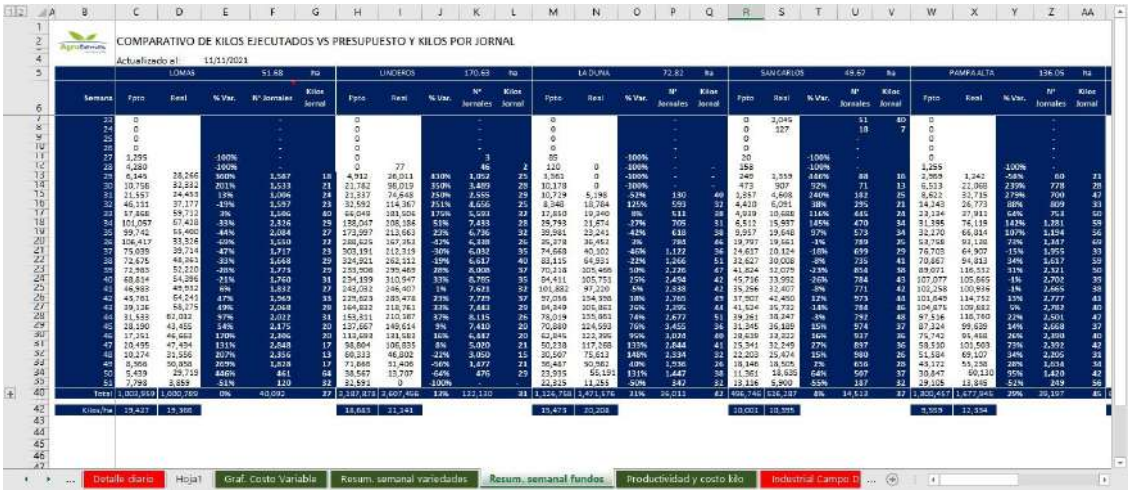
4.2.4. RESULTADO 4 vs OE4: IMPLEMENTAR REDES BAYESIANAS PARA PREDECIR EL CUMPLIMIENTO DE LA CANTIDAD DE ARÁNDANOS COSECHADOS CON BASE EN CARACTERÍSTICAS DEL DETALLE DIARIO HACIENDO USO DEL SOFTWARE SPSS MODELER.

Para realizar la implementación de las redes bayesianas se optó por usar el software SPSS Modeler en su versión 18.0, y los pasos a seguir fueron los siguientes:

1. Del consolidado general, se tuvo que seleccionar solo aquellas variables que iban hacer usadas, las cuales fueron las siguientes:
   - Fundo.
   - Variedad cosechada.
   - Horas trabajadas.
   - Kilos cosechados.
   - Nro de cosechadores.

2. Una vez identificado los atributos, tuvimos que hacer uso de una nueva columna que procedía del Excel original (tablas dinámicas) que manejaba la empresa; ahí se encontraba el detalle del cumplimiento diario de la cosecha ordenado por semana y este se veía como se muestra en la siguiente imagen.

Imagen 108: Detalle del cumplimiento en porcentaje de la cosecha de arándanos



Esta información se mostraba por semana y en porcentaje, por lo cual hicimos un preprocesamiento en Python para unir esta información con el dataset que usaremos (también preprocesado en el lenguaje de programación descrito anteriormente) y que usa las variables mencionadas en el punto 1.

3. Llegado a este punto para el preprocesamiento de las variables a considerar, ya se tenía a estas categorizadas, por tanto, estos datos que han sido trabajados en los objetivos anteriores solamente pasaron a formar grupo con Fundo y Variedad cosechada que ya son columnas nominales, no obstante, el detalle se cada columna queda de la siguiente manera:

   - **Horas trabajadas:** La columna fue renombrar a **Jornada** y se divide en 3 tipos de clases:
     - Mínima: si la jornada fue menor o igual a 7 horas.
     - Normal: si la jornada fue mayor a 7 horas y menor a 9.5 horas.
     - Máxima: si la jornada fue mayor o igual a 9.5 horas.

   - **Kilos cosechados:** La columna fue renombrada a **ValoracionCosecha** y se compone de 3 tipos de clases:
     - Mala: si los kilos cosechados fueron menor o igual a 250 kilos.
     - Normal: si los kilos cosechados fueron mayores a 250 y menor a 800 kilos.
     - Buena: si los kilos cosechados fueron mayores a 800 kilos.

- **Nro de cosechadores:** La columna fue renombrada a **CosechadoresEncode** y las clases que la conforman son 3 y son las siguientes:
  - Baja MO: si el número de cosechadores fueron menor o igual a 8.
  - Regular MO: si el número de cosechadores fueron mayor a 8 y menor a 22.
  - Buena MO: si el número de cosechadores fue mayor o igual a 22.

  Tener en cuenta que MO se refiere mano de obra.

- **Fundo:** Esta columna no se categorizó ni renombró por otra, ya que contiene a clases, mismas que son:
  - Linderos 1.
  - Linderos 2.
  - Lomas.
  - Angostura.
  - Pampa alta.
  - La quebrada.
  - San Carlos.
  - Escalante.
  - La duna.

  Tener en cuenta que, para la clasificación del cumplimiento de la cosecha de arándanos, Linderos 1 y Linderos 2 se tomaron como un mismo fundo, es decir, solamente se le llamo “Linderos”.

- **Variedad de cosecha:** Esta variable tampoco se modificó ya que es columna nominal y se categoriza de la siguiente forma:
  - Biloxi.
  - Atlas.
  - Sekoya Beauty.
  - Sekoya Pop.
  - Ventura.
  - Bianca.
  - Jupiter.

- **CumplimientoKgCosecha:** Esta es la columna que añadimos y explicamos en el punto 2, que en síntesis vendría hacer **KgCosechados Presupuestados vs KgCosechados Real**; esta columna tiene valores numéricos y por tanto también tuvimos que categorizarla, de tal modo que quedó así:
  - Obj. Cosecha incumplido: si el objetivo de la cosecha tiene un progreso de menos del 100%.

- Obj. Cosecha cumplido: si el objetivo de la cosecha tiene un progreso mayor o igual a 100% y menor a 150%.
- Obj. Cosecha Sobresaliente: si el objetivo de la cosecha tiene un progreso mayor o igual a 150%.

4. La transformación de variables numéricas a nominales solo tuvo que ser realizada en la columna denominada **CumplimientoKgCosecha** y esto corresponde a un preprocesamiento que se realizó en Python. Al final se consideró la data filtrada para tomar en cuenta los objetivos cumplidos y sobresalientes.

Imagen 109: Preprocesamiento CumplimientoKgCosecha para OE4

```
[ ]  1 df_final.loc[df_final['KgPPTO vs KgCosechadosReal'] < 100, 'CumplimientoKgCosecha'] = "Obj. Cosecha incumplido"
     2 df_final.loc[(df_final['KgPPTO vs KgCosechadosReal'] >=100) & (df_final['KgPPTO vs KgCosechadosReal'] < 150), 'CumplimientoKgCosecha'] = "Obj. Cosecha cumplido"
     3 df_final.loc[df_final['KgPPTO vs KgCosechadosReal'] >= 150, 'CumplimientoKgCosecha'] = "Obj. Cosecha Sobresaliente"
```

```
1 dataframe
```

	Fundo	VariedadCosechada	Jornada	ValoracionCosecha	CosechadoresEncode	CumplimientoKgCosecha
0	Linderos 2	Biloxi	Maxima	Normal	Regular MO	Obj.Cosecha Sobresaliente
1	Linderos 2	Biloxi	Maxima	Buena	Buena MO	Obj.Cosecha Sobresaliente
2	Linderos 2	Biloxi	Maxima	Normal	Buena MO	Obj.Cosecha Sobresaliente
3	Linderos 2	Biloxi	Maxima	Normal	Regular MO	Obj.Cosecha Sobresaliente
4	Lomas	Biloxi	Minima	Normal	Buena MO	Obj.Cosecha Sobresaliente
...	...	...	...	...	...	...
17962	Escalante	Biloxi	Minima	Mala	Baja MO	Obj.Cosecha Incumplido
17963	Escalante	Ventura	Minima	Normal	Regular MO	Obj.Cosecha Incumplido
17964	San Carlos	Ventura	Minima	Normal	Regular MO	Obj.Cosecha Incumplido
17965	La Duna	Ventura	Minima	Normal	Buena MO	Obj.Cosecha Incumplido
17966	Pampa alta	Ventura	Minima	Normal	Regular MO	Obj.Cosecha Incumplido

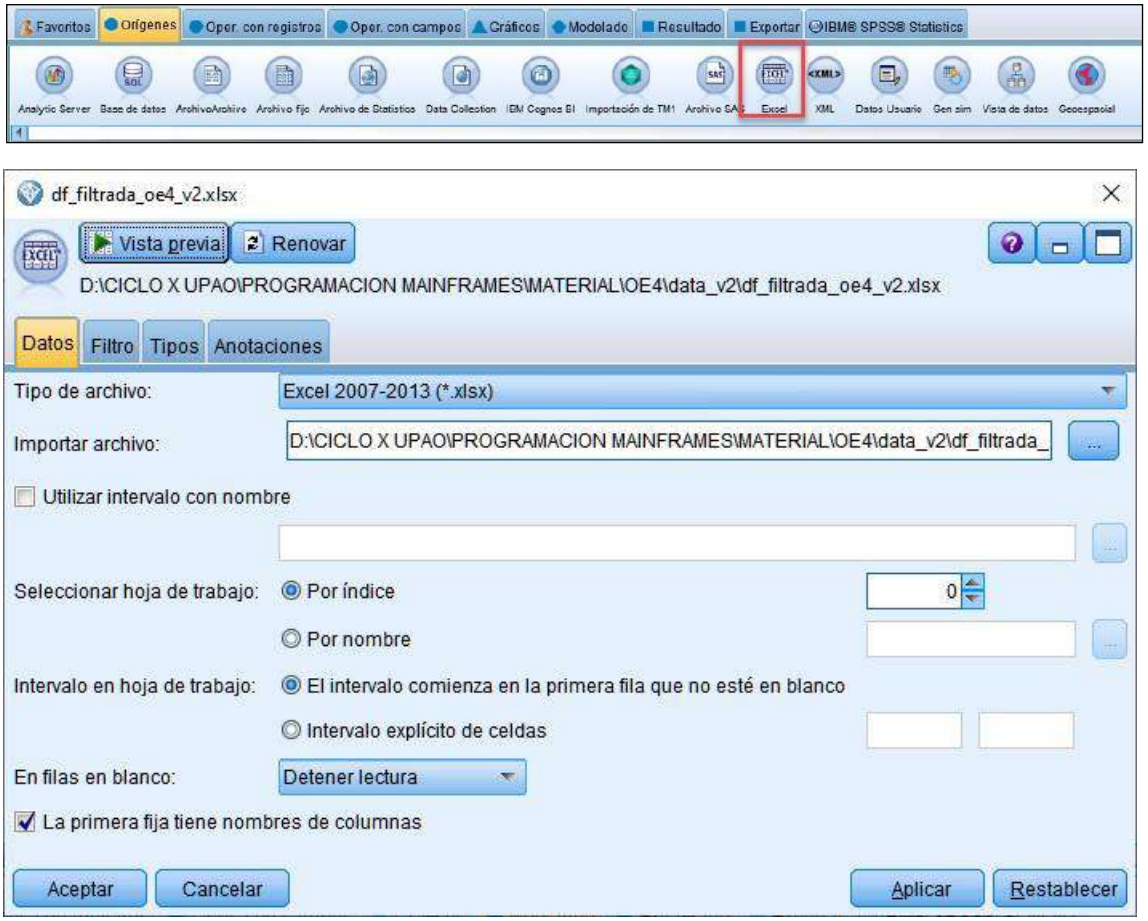
17967 rows × 6 columns

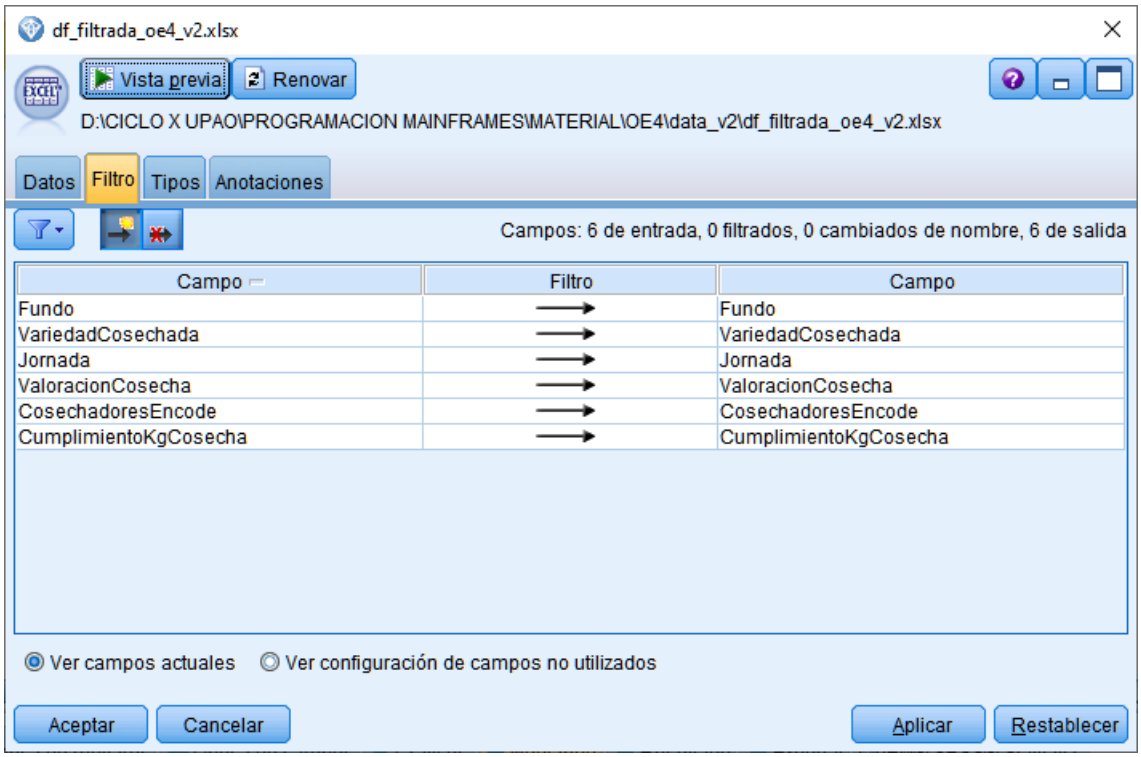
```
1 df_filtrada_oe4_v2=df_final[(df_final['CumplimientoKgCosecha']=='Cosecha completa') | (df_final['CumplimientoKgCosecha']=='Cosecha sobresaliente')]
```

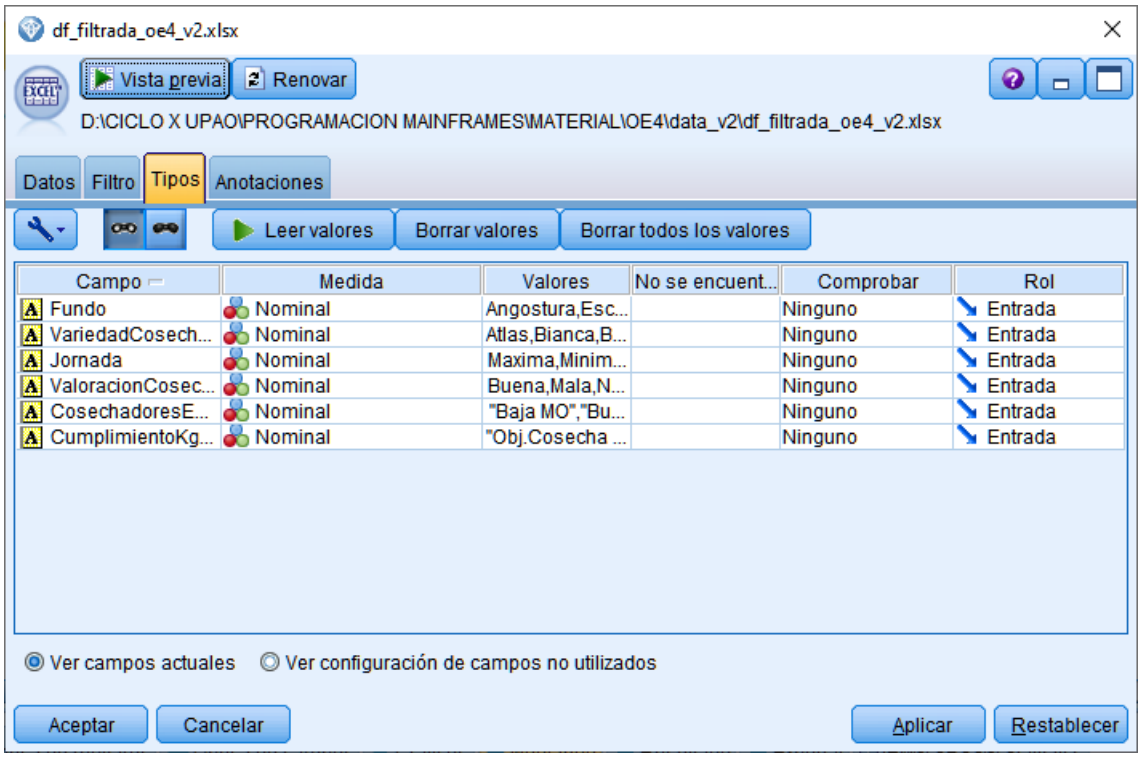
5. A continuación, implementamos la red bayesiana (BayesNet en inglés) haciendo uso del software SPSS Modeler en su versión 18.0, para cual se siguen los pasos mostrados en las siguientes imágenes.

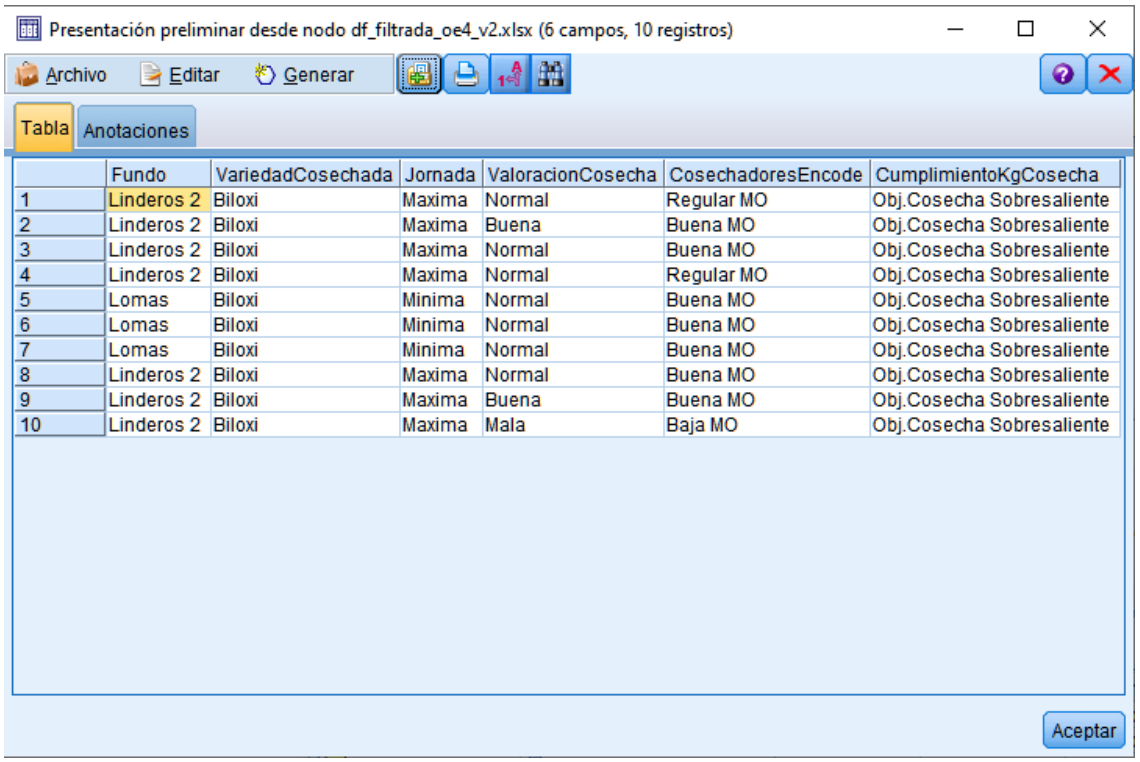
5.1. Iniciamos las configuraciones iniciales en el software: subimos el archivo Excel filtrado con solo las características a usar.

Imagen 110: Importar Excel y configurar la data filtrada en SPSS Modeler




df_filtrada_oe4_v2.xlsx
Vista previa
Renovar
D:\CICLO X UPAO\PROGRAMACION MAINFRAMES\MATERIAL\OE4\data_v2\df_filtrada_oe4_v2.xlsx
Datos
Filtro
Tipos
Anotaciones
Campos: 6 de entrada, 0 filtrados, 0 cambiados de nombre, 6 de salida
Campo
Filtro
Campo
Fundo
VariedadCosechada
Jornada
ValoracionCosecha
CosechadoresEncode
CumplimientoKgCosecha
Ver campos actuales
Ver configuración de campos no utilizados
Aceptar
Cancelar
Aplicar
Restablecer

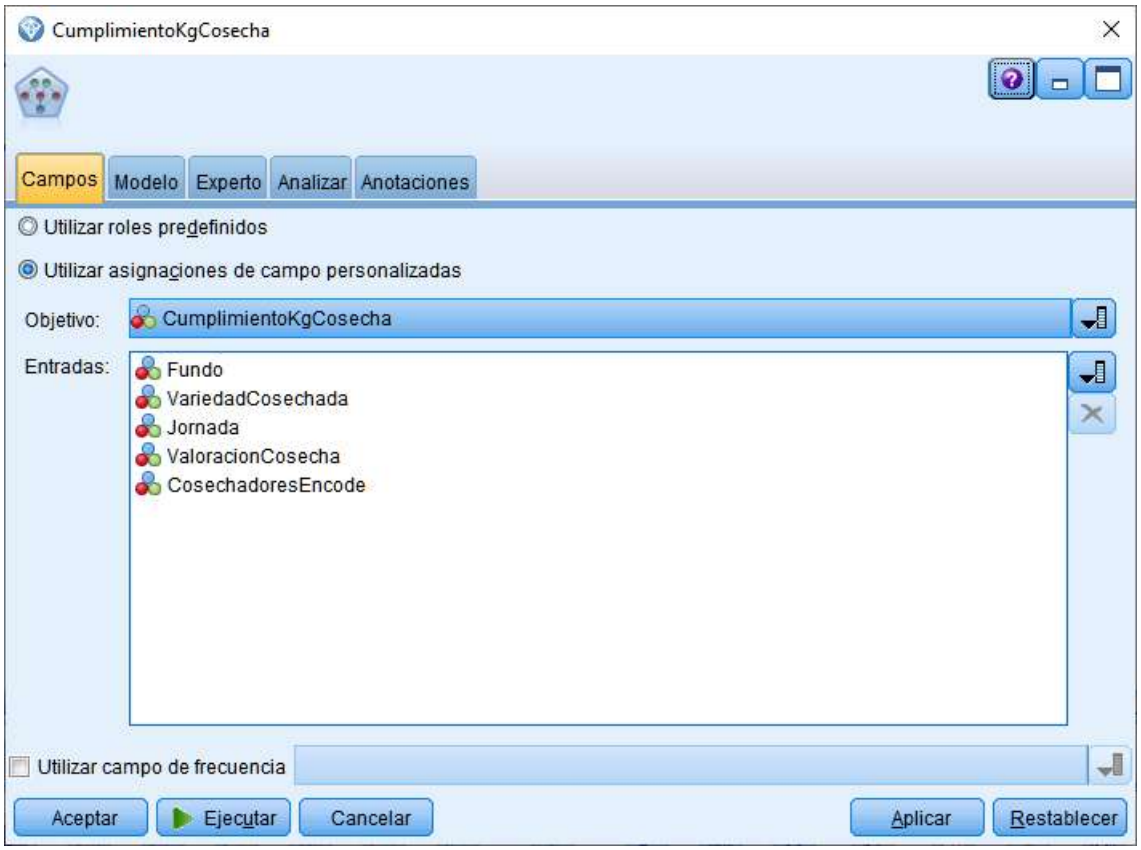

df_filtrada_oe4_v2.xlsx
Vista previa
Renovar
D:\CICLO X UPAO\PROGRAMACION MAINFRAMES\MATERIAL\OE4\data_v2\df_filtrada_oe4_v2.xlsx
Datos
Filtro
Tipos
Anotaciones
Leer valores
Borrar valores
Borrar todos los valores
Campo
Medida
Valores
No se encuent...
Comprobar
Rol
Fundo
Nominal
Angostura,Esc...
Ninguno
Entrada
VariedadCosech...
Atlas,Bianca,B...
Jornada
Maxima,Minim...
ValoracionCosec...
Buena,Mala,N...
CosechadoresE...
"Baja MO","Bu...
CumplimientoKg...
"Obj.Cosecha ...
Ver campos actuales
Ver configuración de campos no utilizados
Aceptar
Cancelar
Aplicar
Restablecer



	Fundo	VariedadCosechada	Jornada	ValoracionCosecha	CosechadoresEncode	CumplimientoKgCosecha
1	Linderos 2	Biloxi	Maxima	Normal	Regular MO	Obj.Cosecha Sobresaliente
2	Linderos 2	Biloxi	Maxima	Buena	Buena MO	Obj.Cosecha Sobresaliente
3	Linderos 2	Biloxi	Maxima	Normal	Buena MO	Obj.Cosecha Sobresaliente
4	Linderos 2	Biloxi	Maxima	Normal	Regular MO	Obj.Cosecha Sobresaliente
5	Lomas	Biloxi	Minima	Normal	Buena MO	Obj.Cosecha Sobresaliente
6	Lomas	Biloxi	Minima	Normal	Buena MO	Obj.Cosecha Sobresaliente
7	Lomas	Biloxi	Minima	Normal	Buena MO	Obj.Cosecha Sobresaliente
8	Linderos 2	Biloxi	Maxima	Normal	Buena MO	Obj.Cosecha Sobresaliente
9	Linderos 2	Biloxi	Maxima	Buena	Buena MO	Obj.Cosecha Sobresaliente
10	Linderos 2	Biloxi	Maxima	Mala	Baja MO	Obj.Cosecha Sobresaliente

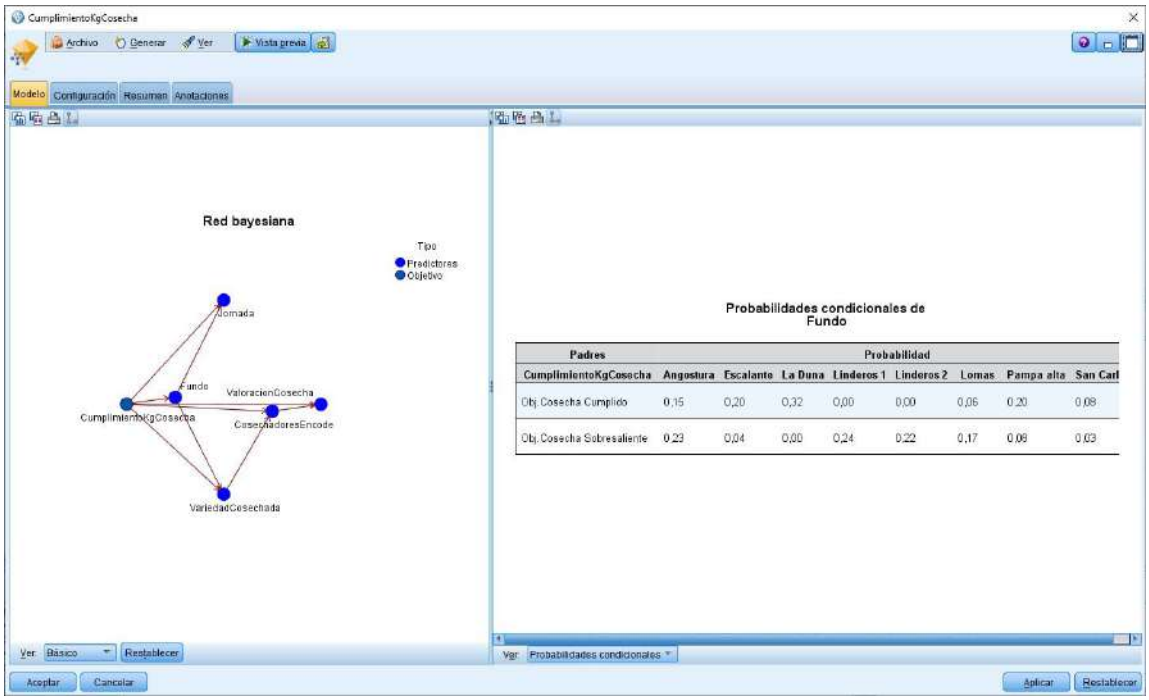
5.2. Configuramos la red bayesiana, la ejecutamos y creamos el modelo.

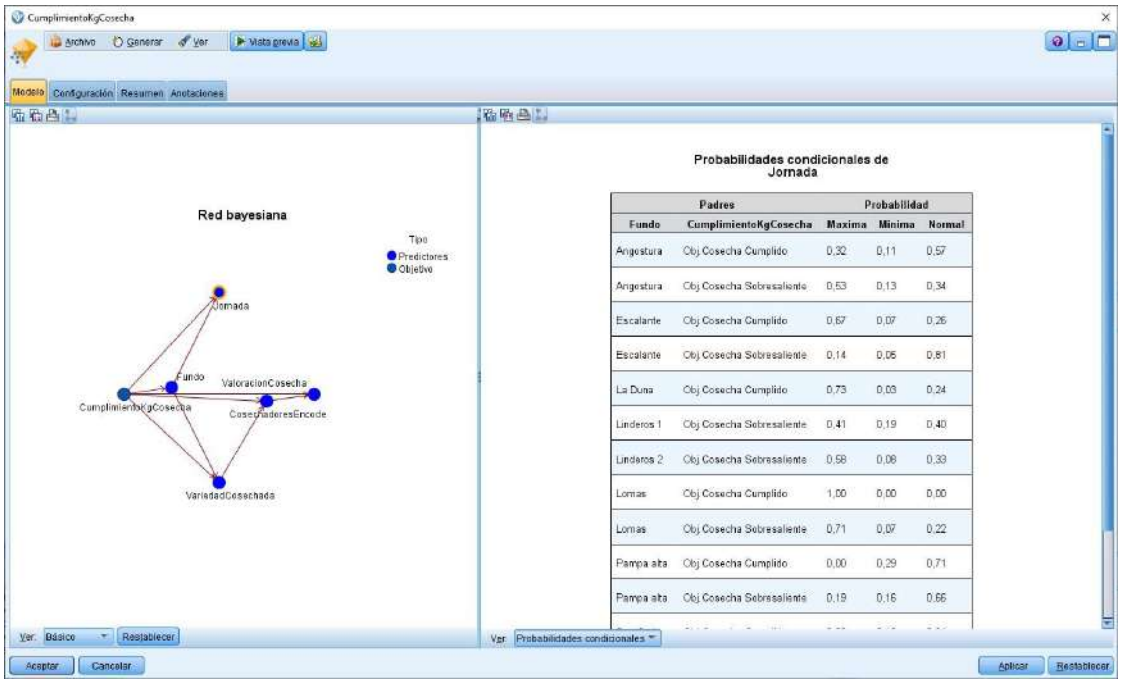
Imagen 111: Configuración del modelo bayesiano en SPSS Modeler

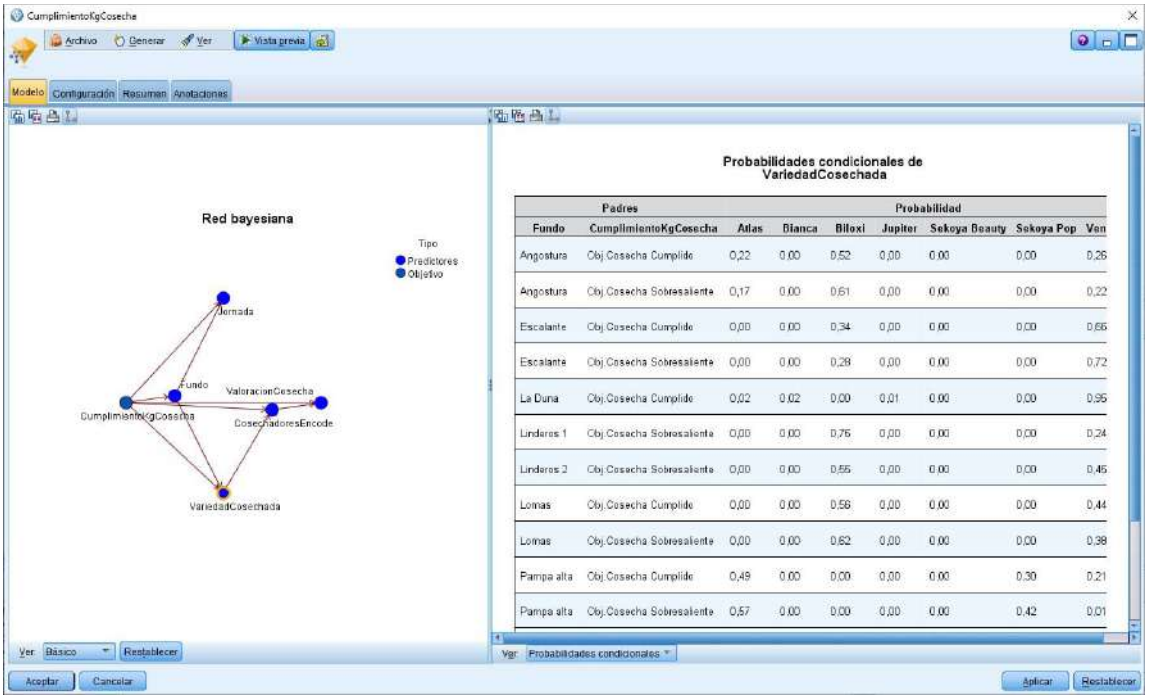


5.3. Analizamos los resultados previos.

Imagen 112: Resultados probabilísticos de la red bayesiana en SPSS Modeler

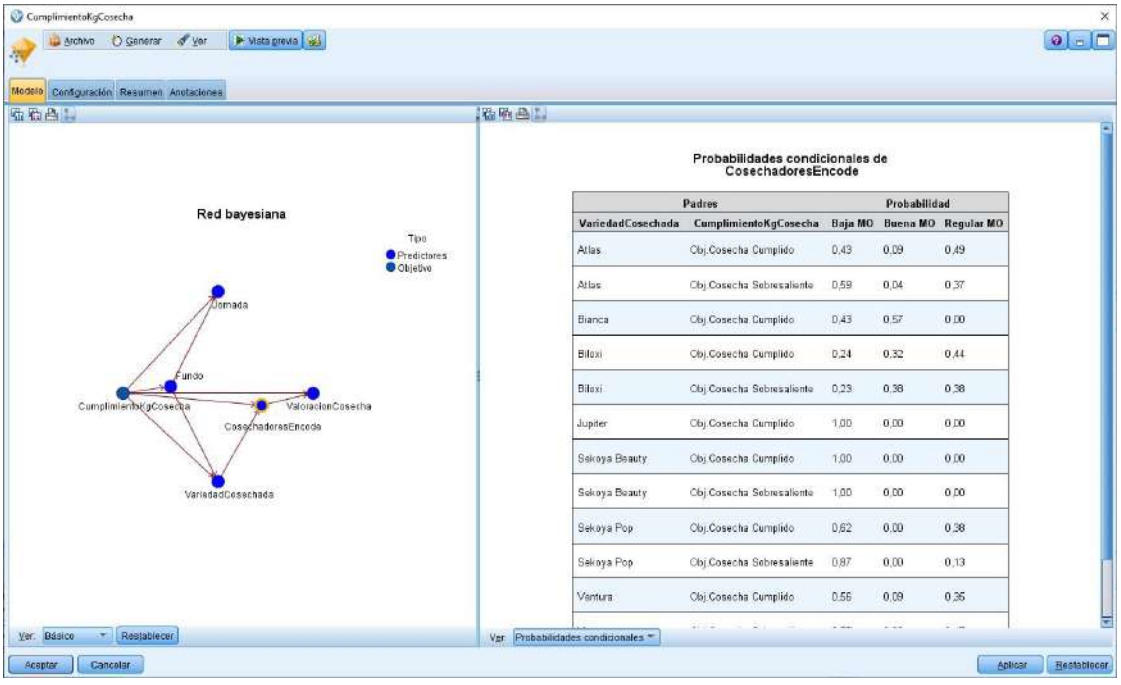






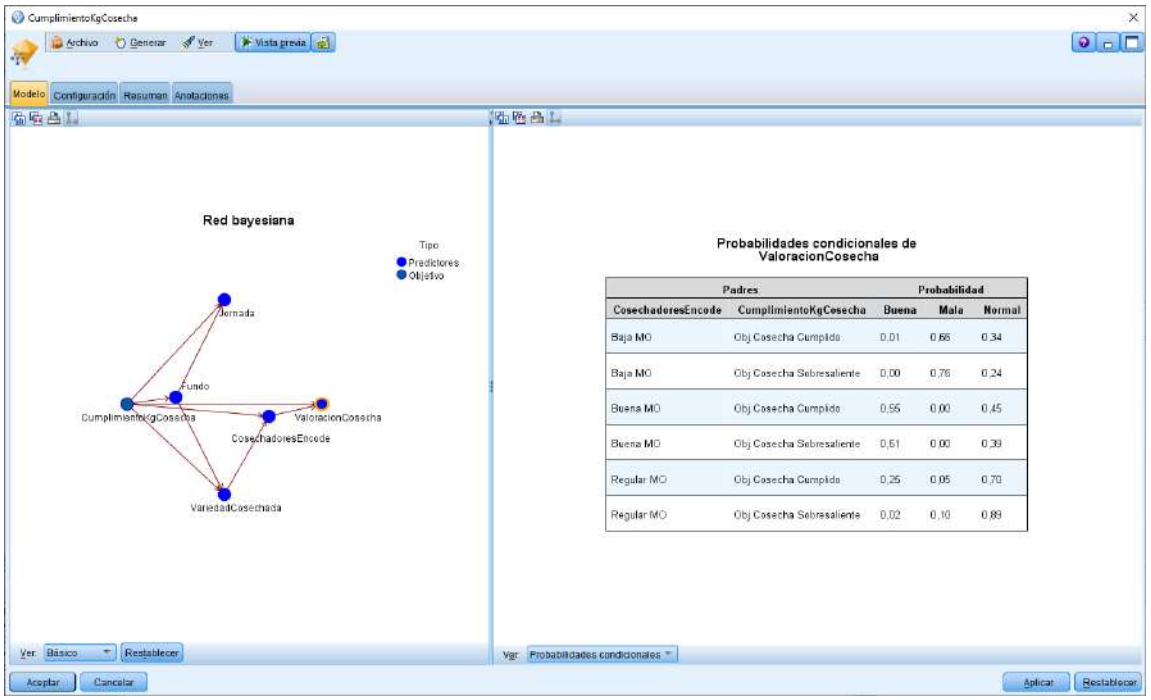
Probabilidades condicionales de VariedadCosechada

Padres		Probabilidad						
Fundo	CumplimientoKgCosecha	Atlas	Bianca	Biloxi	Jupiter	Sekoya Beauty	Sekoya Pop	Ven
Angostura	Obj.Cosecha Cumplido	0,22	0,00	0,52	0,00	0,00	0,00	0,26
Angostura	Obj.Cosecha Sobresaliente	0,17	0,00	0,61	0,00	0,00	0,00	0,22
Escalante	Obj.Cosecha Cumplido	0,00	0,00	0,34	0,00	0,00	0,00	0,66
Escalante	Obj.Cosecha Sobresaliente	0,00	0,00	0,28	0,00	0,00	0,00	0,72
La Duna	Obj.Cosecha Cumplido	0,02	0,02	0,00	0,01	0,00	0,00	0,95
Linderos 1	Obj.Cosecha Sobresaliente	0,00	0,00	0,76	0,00	0,00	0,00	0,24
Linderos 2	Obj.Cosecha Sobresaliente	0,00	0,00	0,55	0,00	0,00	0,00	0,45
Lomas	Obj.Cosecha Cumplido	0,00	0,00	0,56	0,00	0,00	0,00	0,44
Lomas	Obj.Cosecha Sobresaliente	0,00	0,00	0,62	0,00	0,00	0,00	0,38
Pampa alta	Obj.Cosecha Cumplido	0,49	0,00	0,00	0,00	0,00	0,30	0,21
Pampa alta	Obj.Cosecha Sobresaliente	0,57	0,00	0,00	0,00	0,00	0,42	0,01



Probabilidades condicionales de CosechadoresEncode

Padres		Probabilidad		
VariedadCosechada	CumplimientoKgCosecha	Baja MO	Buena MO	Regular MO
Atlas	Obj.Cosecha Cumplido	0,43	0,09	0,49
Atlas	Obj.Cosecha Sobresaliente	0,59	0,04	0,37
Bianca	Obj.Cosecha Cumplido	0,43	0,57	0,00
Biloxi	Obj.Cosecha Cumplido	0,24	0,32	0,44
Biloxi	Obj.Cosecha Sobresaliente	0,23	0,38	0,38
Jupiter	Obj.Cosecha Cumplido	1,00	0,00	0,00
Sekoya Beauty	Obj.Cosecha Cumplido	1,00	0,00	0,00
Sekoya Beauty	Obj.Cosecha Sobresaliente	1,00	0,00	0,00
Sekoya Pop	Obj.Cosecha Cumplido	0,62	0,00	0,38
Sekoya Pop	Obj.Cosecha Sobresaliente	0,87	0,00	0,13
Ventura	Obj.Cosecha Cumplido	0,56	0,09	0,35



5.4. Obtuvimos la matriz de coincidencia y estadísticas.

Tabla 25: Matriz de coincidencia de la red bayesiana

	Obj. Cosecha Cumplido	***Obj. Cosecha Sobresaliente***
Obj. Cosecha Cumplido	933	308
Obj. Cosecha Sobresaliente	148	1,812

Tabla 26: Valores de confianza de la red bayesiana

Rango	0, 513 – 1,0
Media para correctos	0,888
Media para incorrectos	0,685
Siempre correctos por encima de	0,972 (45,58% de casos)
Siempre incorrectos por debajo de	0,513 (0% de casos)
90,12% Precisión por encima de	0,639
2,0 veces correctos por encima de	0,665 (93,02% de casos)

Los resultados de este objetivo nos brindan algunas de los siguientes detalles:

- Con respecto a fundos y el cumplimiento de la cosecha:
    - Se predice en un **0%** de probabilidades de que en **La Duna** no habrá cosecha sobresaliente.
    - Se predice en un **0%** de probabilidades de que en **Linderos 1** y en **Linderos 2** no habrá una cosecha cumplida.
    - Una **solución propuesta** para los casos anteriores sería valorar la posibilidad de distribuir las siembras en otros fundos y observar en detalle las características de estos.
- Con respecto a la jornada y fundos:
    - Se predice en un **100%** de probabilidades de que si el fundo es **Lomas** habrá una **cosechada cumplida** siempre y cuando la jornada sea **máxima**.
    - Se predice en un **0%** de probabilidades de que si el fundo es **Pampa Alta** no habrá **cosecha cumplida** incluso si la jornada es **máxima**.
    - Una **solución propuesta** para los casos anteriores serían la consideración de estos dos fundos y la gestión de la jornada laboral para los cosechadores.
- Con respecto a los tipos de arándanos (variedad cosechad) y fundos:
    - Se predice en un **95%** de probabilidades de que si el fundo es **La Duna** y se cosecha **Ventura** como tipo de arándano entonces se obtendrá una **cosecha cumplida**.
    - En el fundo **La Duna** se predice en un **0%** de probabilidades de que los tipos de arándanos **Biloxi**, **Sekoya Beauty**, **Sekoya Pop**, además de que **Atlas**, **Bianca** y **Jupiter** obtienen un **2%** de probabilidades en las dos primeras y **1%** en la restante que se obtendrá una **cosecha cumplida**.
    - Una **solución propuesta** sería gestionar de forma adecuada las variedades cosechadas en el fundo La Duna y tomar en cuenta las prioridades de cosecha.
- Con respecto a la mano de obra (cosechadores encode) y la valoración de la cosecha:
    - Se predice en un **0%** de probabilidades que si la mano de obra es **baja** y se obtiene una valoración de la cosecha **buena** se obtendrá un objetivo de **cosecha sobresaliente**.
    - Se predice en un **89%** de probabilidades que si la mano de obra es **regular** y se obtiene una valoración de la cosecha **normal** se obtendrá un objetivo de **cosecha sobresaliente**.
    - Una **solución propuesta** sería gestionar la mano de obra tras obtener una valoración de la cosecha no esperada.

4.2.5. RESULTADO 5 vs OE5: VISUALIZACIÓN DE LOS RESULTADOS ESTADÍSTICOS A TRAVÉS DE GRÁFICOS OBTENIDOS POR CADA SOFTWARE.

A continuación, se muestran las visualizaciones resultantes de los modelos aplicados en el presente trabajo ordenados por número de objetivo, excluyendo al OE1:

- **Para el OE2:**

Con la finalidad de que el decisor final pueda observar de forma más estructurada y tomar decisiones con mayor criterio, se realizó una tabulación cruzada por cada variable que se tomó en cuenta en este objetivo y se tomó los datos de prueba **(30% para test)** para los resultados finales.

Imagen 113: Tabulación cruzada Jornada por Sobrecosto

Cross Tabulation of Jornada by Sobrecosto

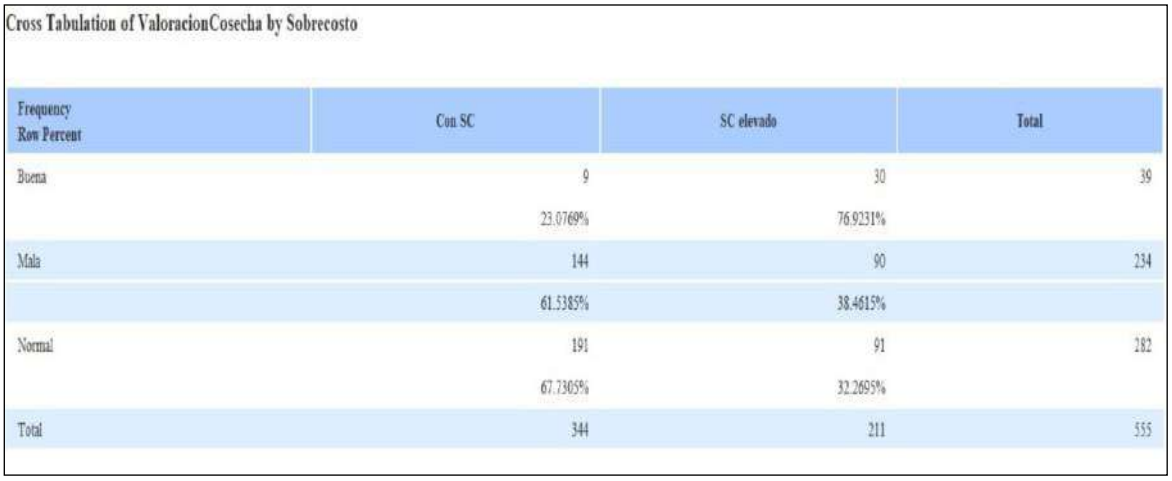
Frequency Row Percent	Con SC	SC elevado	Total
Maxima	268	58	326
	82.2086%	17.7914%	
Minima	1	35	36
	2.7778%	97.2222%	
Normal	75	118	193
	38.8601%	61.1399%	
Total	344	211	555

Imagen 114: Tabulación cruzada Variedad de arándano cosechada por Sobrecosto

Cross Tabulation of VariedadCosechada by Sobrecosto

Frequency Row Percent	Con SC	SC elevado	Total
Atlas	44	19	63
	69.8413%	30.1587%	
Biloxi	62	98	160
	38.75%	61.25%	
Jupiter	1		1
	100%		
Sekoya Beauty	5	6	11
	45.4545%	54.5455%	
Sekoya Pop	18	13	31
	58.0645%	41.9355%	
Ventura	214	75	289
	74.0484%	25.9516%	
Total	344	211	555

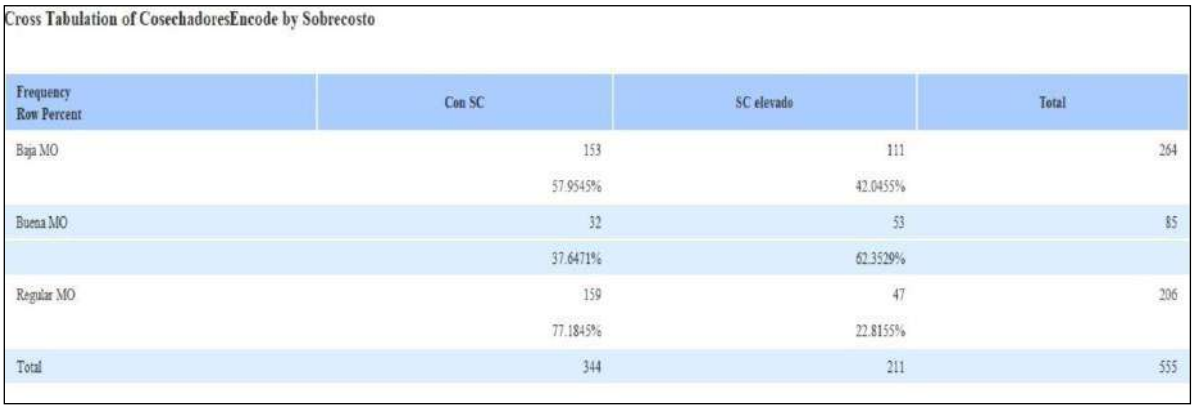
Imagen 115: Tabulación cruzada Valoración de la cosecha por Sobrecosto



Cross Tabulation of ValoracionCosecha by Sobrecosto

Frequency Row Percent	Con SC	SC elevado	Total
Buena	9	30	39
	23.0769%	76.9231%	
Mala	144	90	234
	61.5385%	38.4615%	
Normal	191	91	282
	67.7305%	32.2695%	
Total	344	211	555

Imagen 116: Tabulación cruzada Mano de obra (CosechadoresEncode) por Sobrecosto



Cross Tabulation of CosechadoresEncode by Sobrecosto

Frequency Row Percent	Con SC	SC elevado	Total
Baja MO	153	111	264
	57.9545%	42.0455%	
Buena MO	32	53	85
	37.6471%	62.3529%	
Regular MO	159	47	206
	77.1845%	22.8155%	
Total	344	211	555

- **<u>Para el OE3:</u>**

Se realizaron distintos gráficos en barras asumiendo los valores estadísticos de las reglas de asociación (soporte, confianza y lift) de los resultados del programa de minería de datos Orange apoyado en el software Excel para graficar.

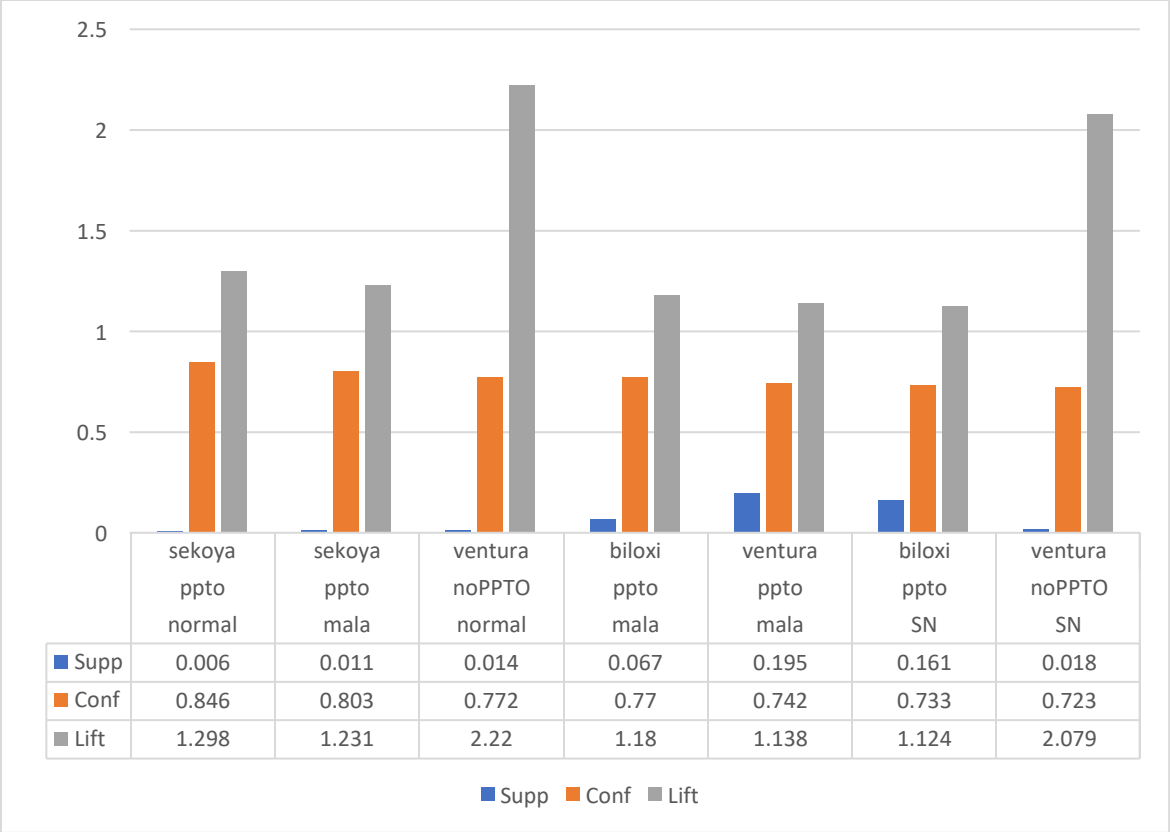
Tabla 27: Reglas de asociación ordenadas horizontalmente

ValoracionCosecha	Costo	VariedadArandano	Supp	Conf	Lift
normal	ppto	sekoya	0.006	0.846	*1.298*
mala	ppto	sekoya	0.011	0.803	1.231
normal	noPPTO	ventura	0.014	0.772	2.22
mala	ppto	biloxi	0.067	0.77	1.18
mala	ppto	ventura	0.195	0.742	1.138
SN	ppto	biloxi	0.161	0.733	1.124
SN	noPPTO	ventura	0.018	0.723	2.079

Fuente: Elaboración propia

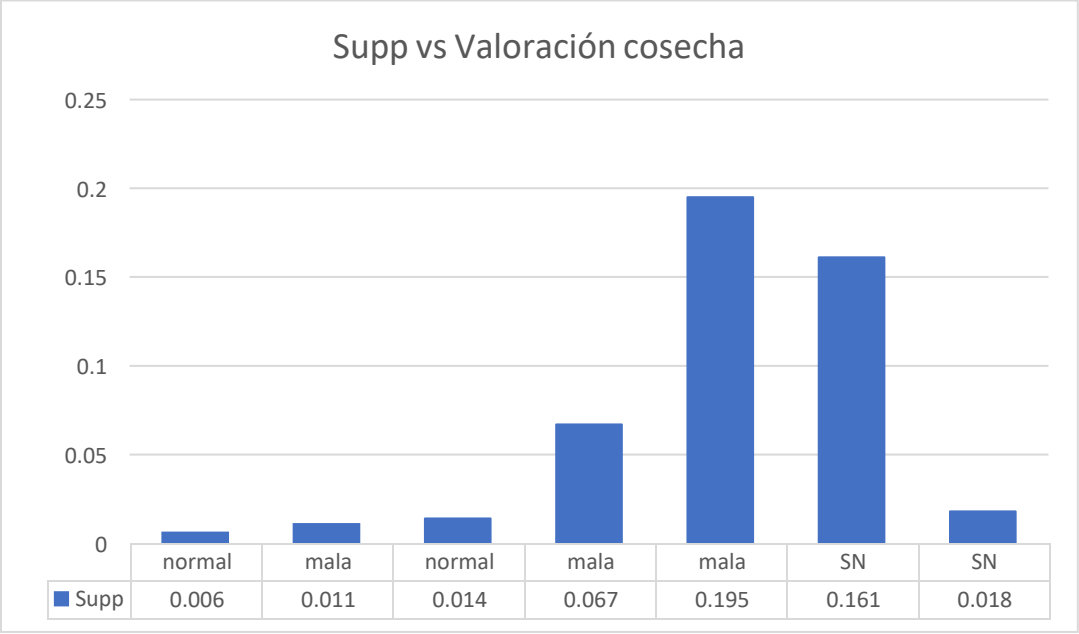
A partir de ello, se dedujeron los siguientes gráficos:

Imagen 117: Gráfico de barras de todas reglas de asociación



	sekoya ppto normal	sekoya ppto mala	ventura noPPTO normal	biloxi ppto mala	ventura ppto mala	biloxi ppto SN	ventura noPPTO SN
Supp	0.006	0.011	0.014	0.067	0.195	0.161	0.018
Conf	0.846	0.803	0.772	0.77	0.742	0.733	0.723
Lift	1.298	1.231	2.22	1.18	1.138	1.124	2.079

Imagen 118: Gráfico de barras soporte vs Valoración cosecha



	normal	mala	normal	mala	mala	SN	SN
Supp	0.006	0.011	0.014	0.067	0.195	0.161	0.018

Imagen 119: Gráfico de barras soporte vs Costo

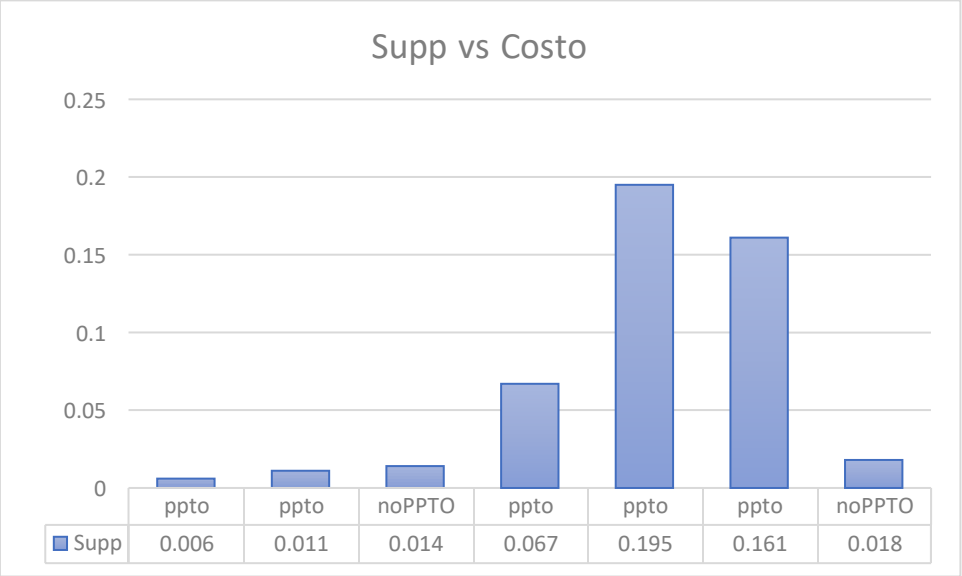


Imagen 120: Gráfico de barras soporte vs Variedad de arándano cosechada

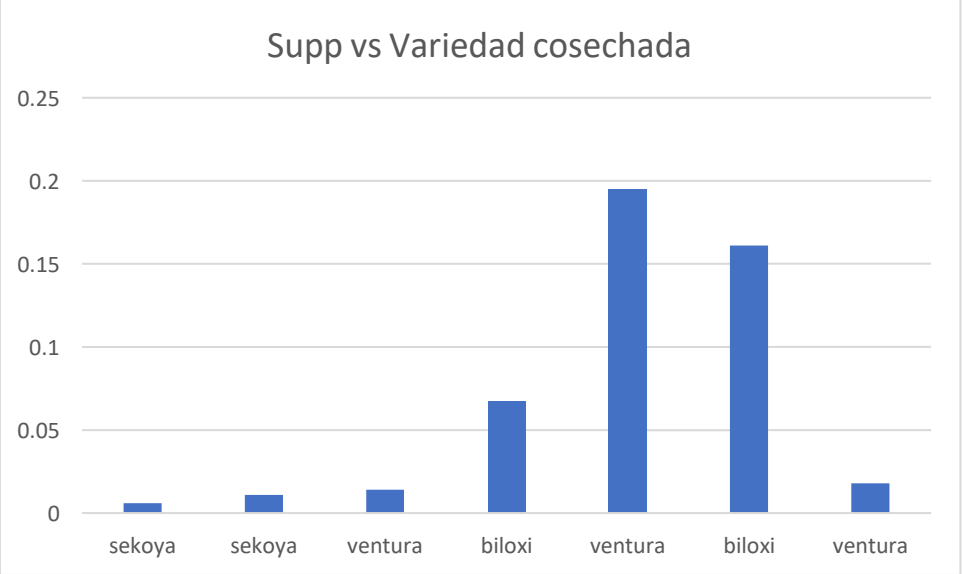


Imagen 121: Gráfico de barras confianza vs Valoración cosecha

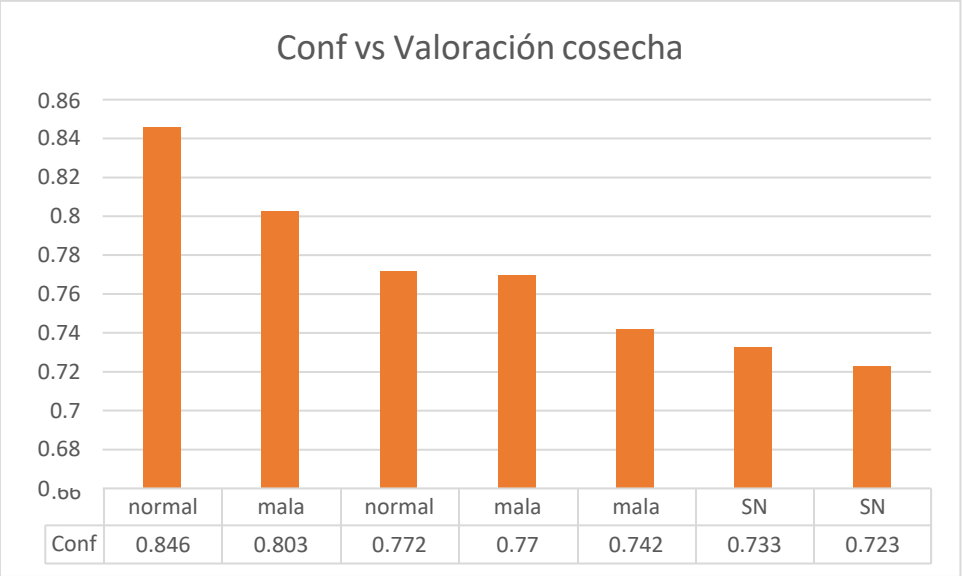


Imagen 122: Gráfico de barras confianza vs Costo

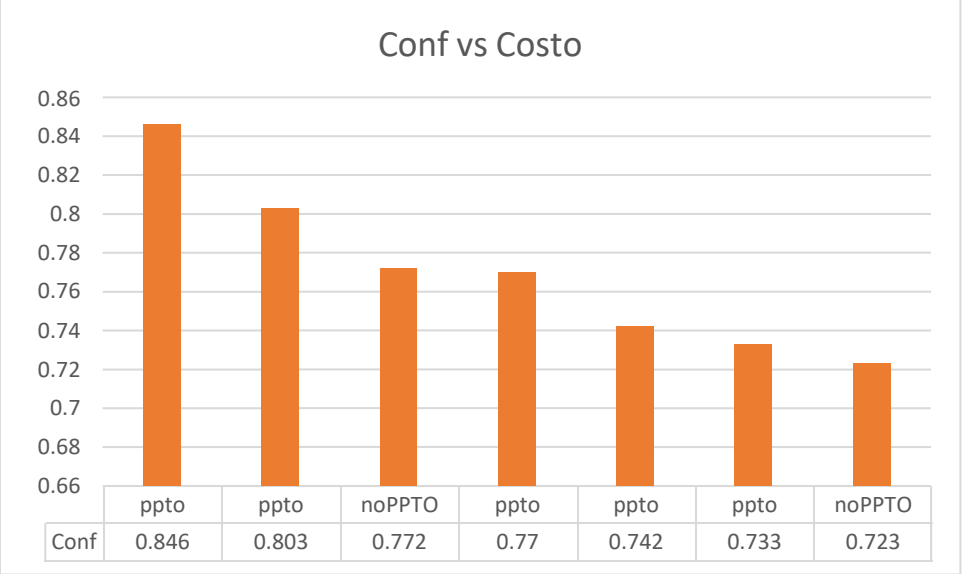


Imagen 123: Gráfico de barras confianza vs Variedad arándano cosechada

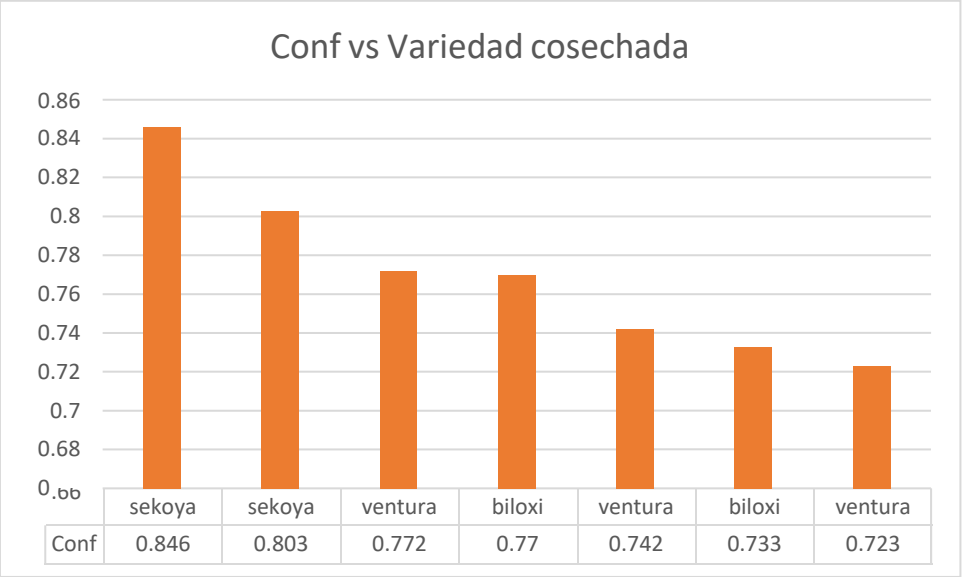


Imagen 124: Gráfico de barras Lift vs Valoración cosecha

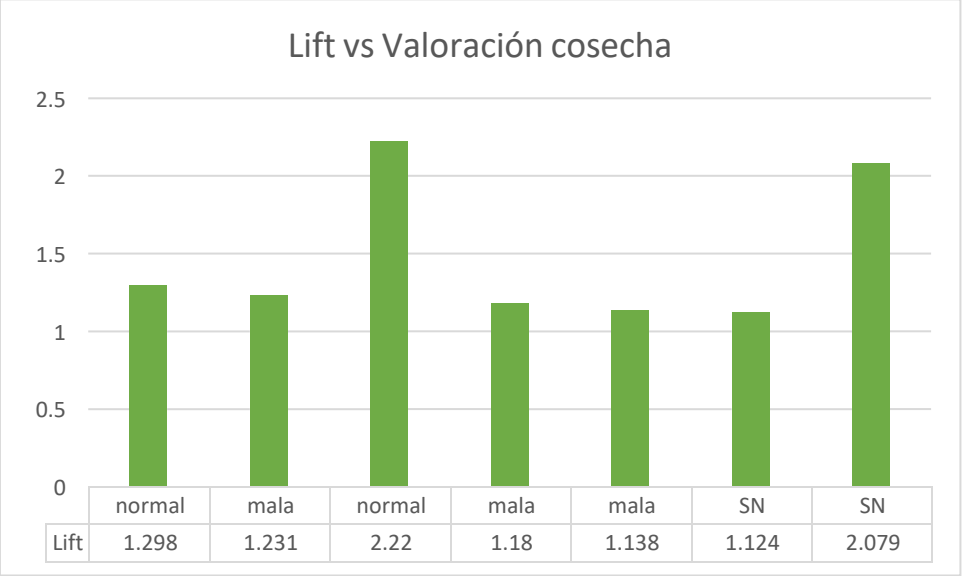


Imagen 125: Gráfico de barras Lift vs Costo

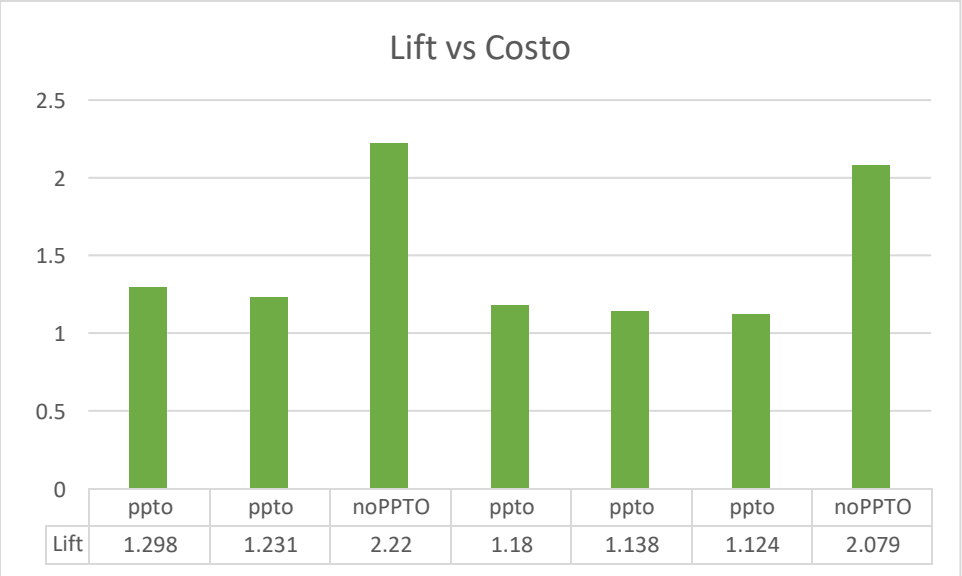
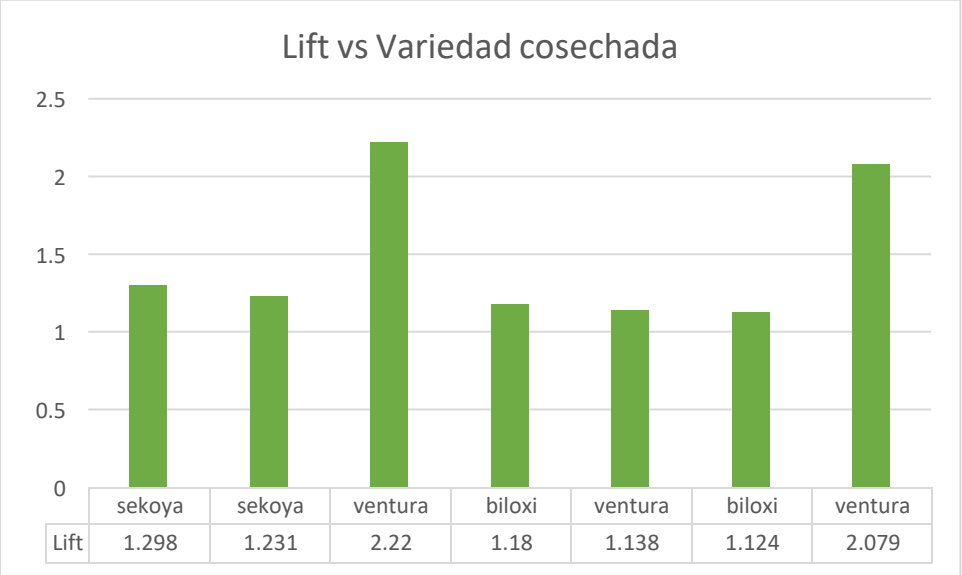


Imagen 126: Gráfico de barras Lift vs Variedad de arándano cosechada



- **Para el OE4:**

Se realizaron gráfico de mapa de calor apoyado por el software SPSS Modeler.

Imagen 127: Mapa de calor CumplimientoKgCosecha vs Jornada laboral

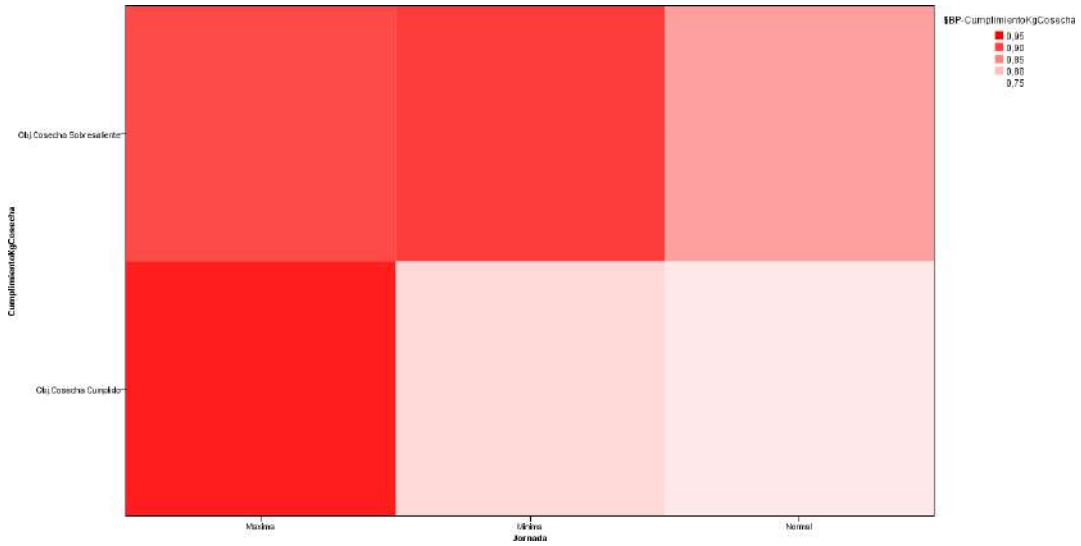


Tabla 28: Resultados del mapa de calor Cumplimiento KgCosecha vs Jornada laboral

Jornada	***CumplimientoKgCosecha***	***$BP-CumplimientoKgCosecha***
Maxima	Obj.Cosecha Sobresaliente	0.8916386078049564
Minima	Obj.Cosecha Sobresaliente	0.902690934382769
Normal	Obj.Cosecha Sobresaliente	0.8251416111957701
Minima	Obj.Cosecha Cumplido	0.7803012034400773
Maxima	Obj.Cosecha Cumplido	0.9274491194319552
Normal	Obj.Cosecha Cumplido	0.7681436792799861

Imagen 128: Mapa de calor CumplimientoKgCosecha vs Variedad de arándano cosechada

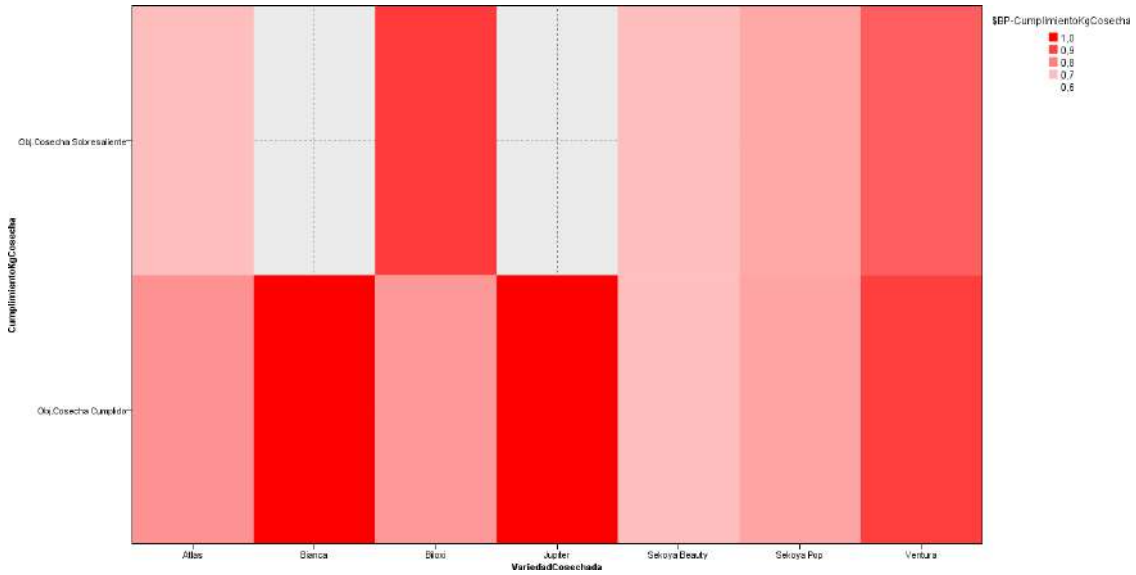


Tabla 29: Resultados del mapa de calor Cumplimiento KgCosecha vs Variedad de arándano cosechada

VariedadCosechada	***CumplimientoKgCosecha***	***$BP-CumplimientoKgCosecha***
Biloxi	Obj.Cosecha Sobresaliente	0.907513718232029
Atlas	Obj.Cosecha Sobresaliente	0.7023123006751333
Ventura	Obj.Cosecha Sobresaliente	0.8521959300720646
Sekoya Pop	Obj.Cosecha Sobresaliente	0.7369300797611573
Sekoya Beauty	Obj.Cosecha Sobresaliente	0.6998243192129499
Atlas	Obj.Cosecha Cumplido	0.7717276507734172
Ventura	Obj.Cosecha Cumplido	0.9027531951404133
Bianca	Obj.Cosecha Cumplido	1.0
Biloxi	Obj.Cosecha Cumplido	0.7579008775258328
Sekoya Beauty	Obj.Cosecha Cumplido	0.6998243192129499
Sekoya Pop	Obj.Cosecha Cumplido	0.7410936184923914
Jupiter	Obj.Cosecha Cumplido	1.0

Imagen 129: Mapa de calor CumplimientoKgCosecha vs Valoración de la cosecha

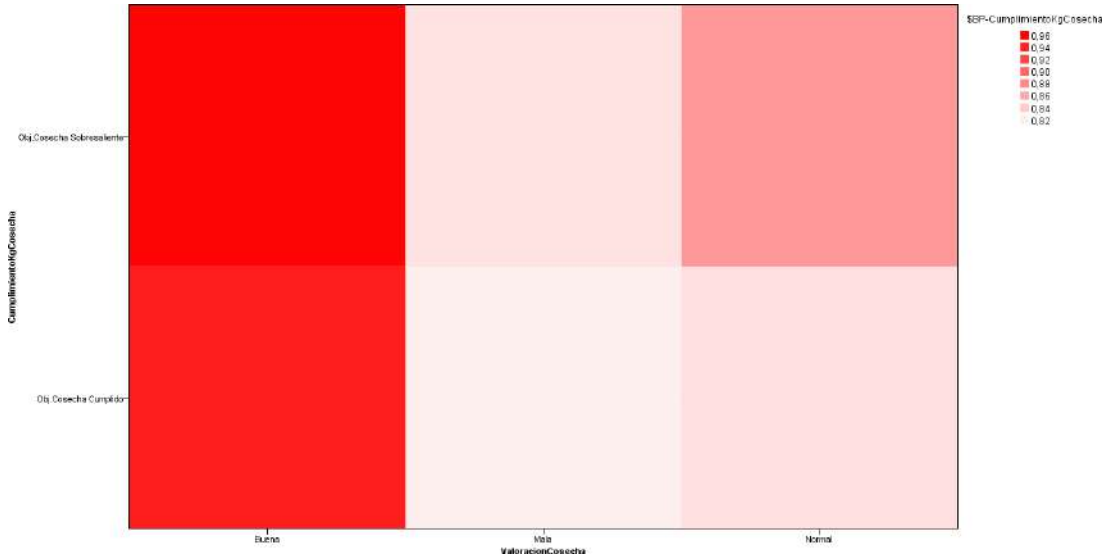


Tabla 30: Resultados del mapa de calor Cumplimiento KgCosecha vs Valoración de la cosecha

ValoracionCosecha	***CumplimientoKgCosecha***	***$BP-CumplimientoKgCosecha***
Normal	Obj.Cosecha Sobresaliente	0.8707650218030185
Buena	Obj.Cosecha Sobresaliente	0.9574914355673336
Mala	Obj.Cosecha Sobresaliente	0.8266723545851316
Mala	Obj.Cosecha Cumplido	0.8186438676405519
Normal	Obj.Cosecha Cumplido	0.8287578802948103
Buena	Obj.Cosecha Cumplido	0.9425226213525023

Imagen 130: Mapa de calor CumplimientoKgCosecha vs Mano de obra (CosechadoresEncode)

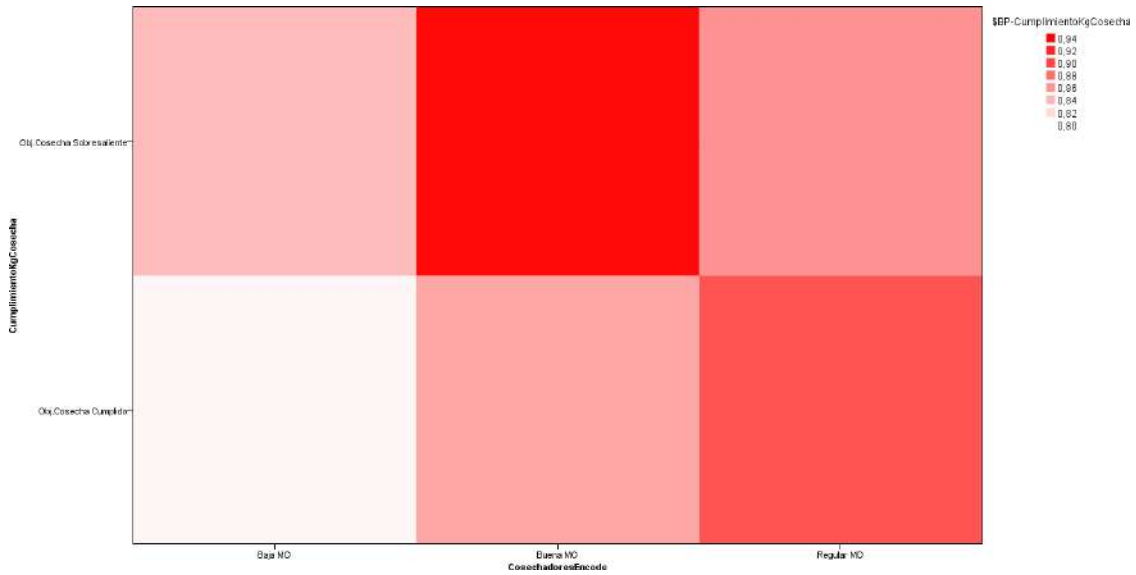


Tabla 31: Resultados del mapa de calor Cumplimiento KgCosecha vs Mano de obra (CosechadoresEncode)

CosechadoresEncode	***CumplimientoKgCosecha***	***$BP-CumplimientoKgCosecha***
Regular MO	Obj.Cosecha Sobresaliente	0.8594822239941173
Buena MO	Obj.Cosecha Sobresaliente	0.9349877727582441
Baja MO	Obj.Cosecha Sobresaliente	0.8367644289773855
Baja MO	Obj.Cosecha Cumplido	0.804621578319183
Regular MO	Obj.Cosecha Cumplido	0.8938506008306856
Buena MO	Obj.Cosecha Cumplido	0.8491129654895937

4.3. DOCIMASIA DE HIPÓTESIS

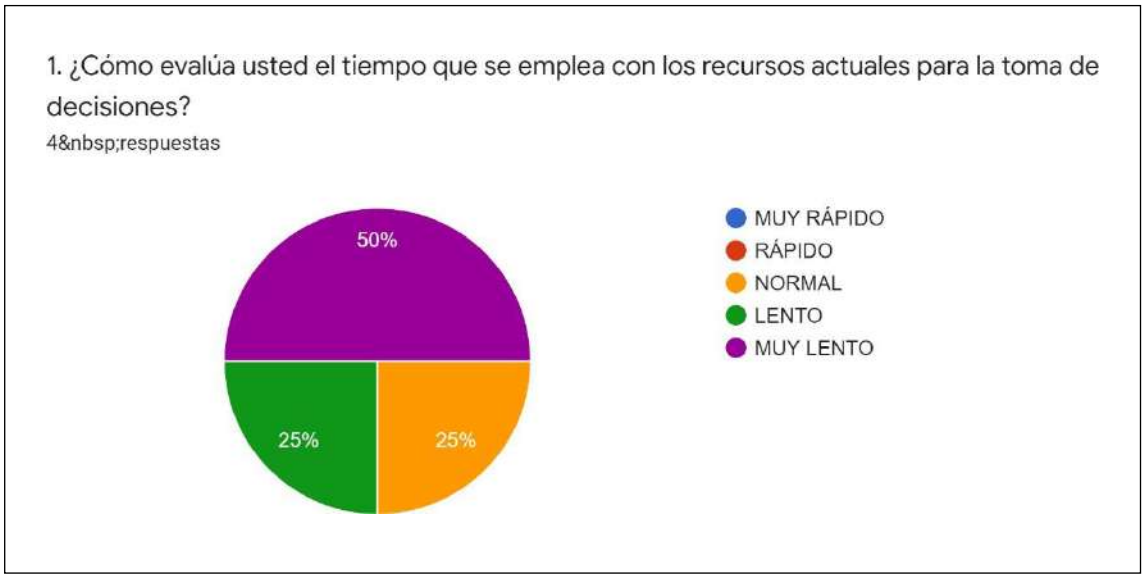
Se realizó una encuesta la cual usó como escala a la de Likert y se les hizo entrega a los tomadores de decisiones antes y después de aplicada la minería de datos para que pudiesen auto evaluarse referente a la satisfacción que perciben de la toma de decisiones sin y con una minería de datos realizada de forma correcta además de hacer uso de varias técnicas con la interpretación gráfica y estadística de cada una de ellas.

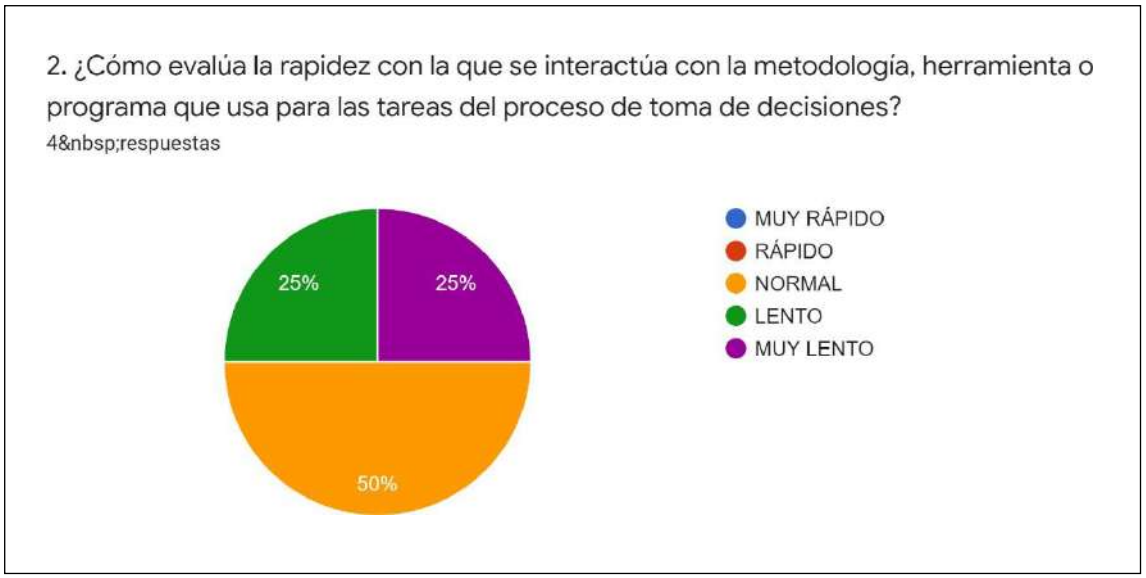
H_1 = La aplicación de la minería de datos y aprendizaje automático **sirve** como herramienta de soporte para la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022.

H_0 = La aplicación de la minería de datos y aprendizaje automático **no sirve** como herramienta de soporte para la toma de decisiones en el área de producción y transporte de arándanos en la empresa Agroberries S.A.C. – La Libertad 2022.

4.3.1. RESULTADOS PRE-MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO

Imagen 131: Resultados de las encuestas pre-minería de datos y aprendizaje automático - Eficacia




2. ¿Cómo evalúa la rapidez con la que se interactúa con la metodología, herramienta o programa que usa para las tareas del proceso de toma de decisiones?
4 respuestas
25%
25%
50%
MUY RÁPIDO
RÁPIDO
NORMAL
LENTO
MUY LENTO

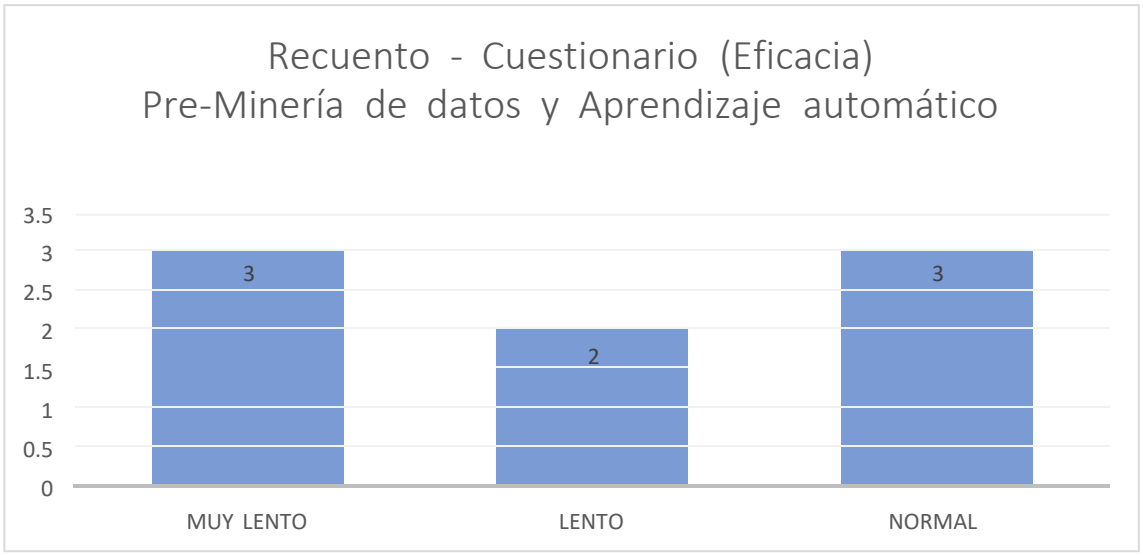
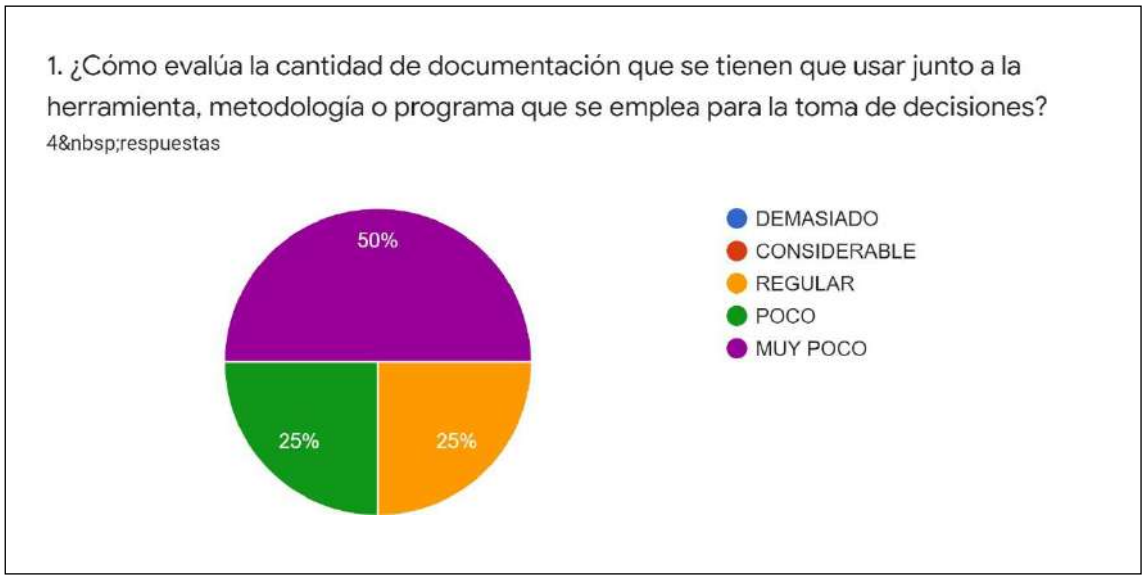
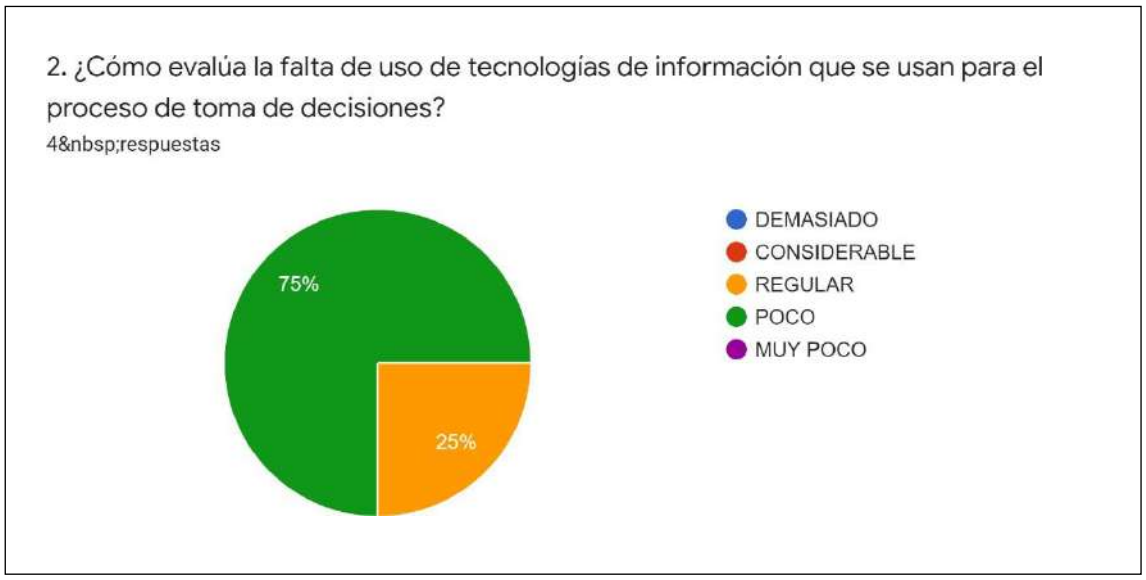

Recuento - Cuestionario (Eficacia)
Pre-Minería de datos y Aprendizaje automático
3.5
3
2.5
2
1.5
1
0.5
0
3
2
3
MUY LENTO
LENTO
NORMAL

Imagen 132: Resultados de las encuestas pre-minería de datos y aprendizaje automático – Eficiencia





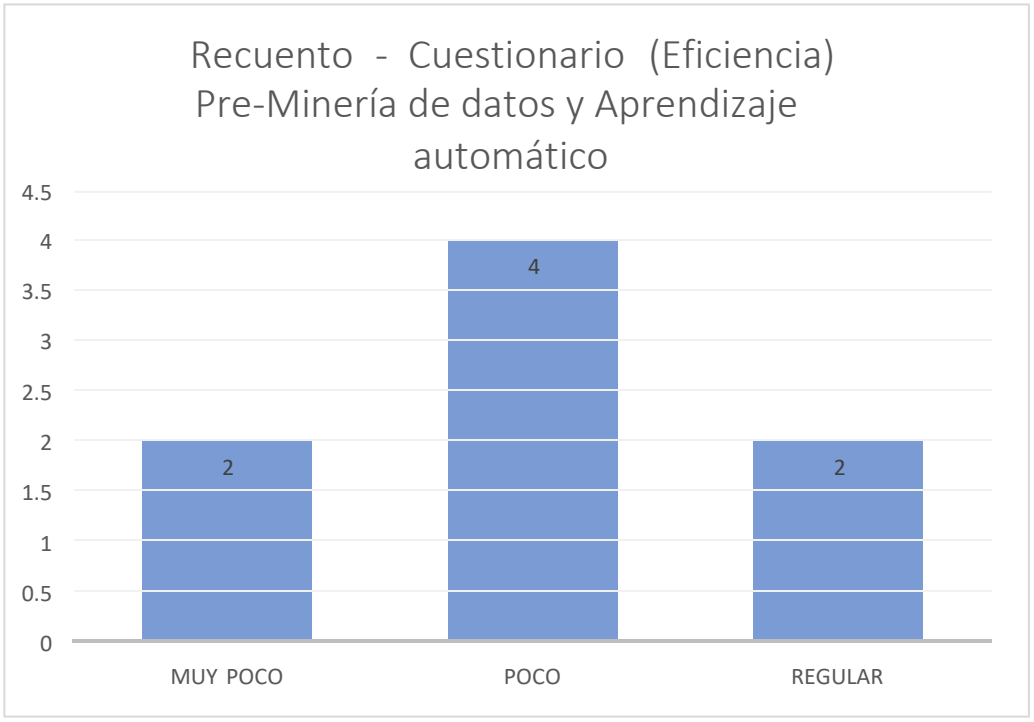
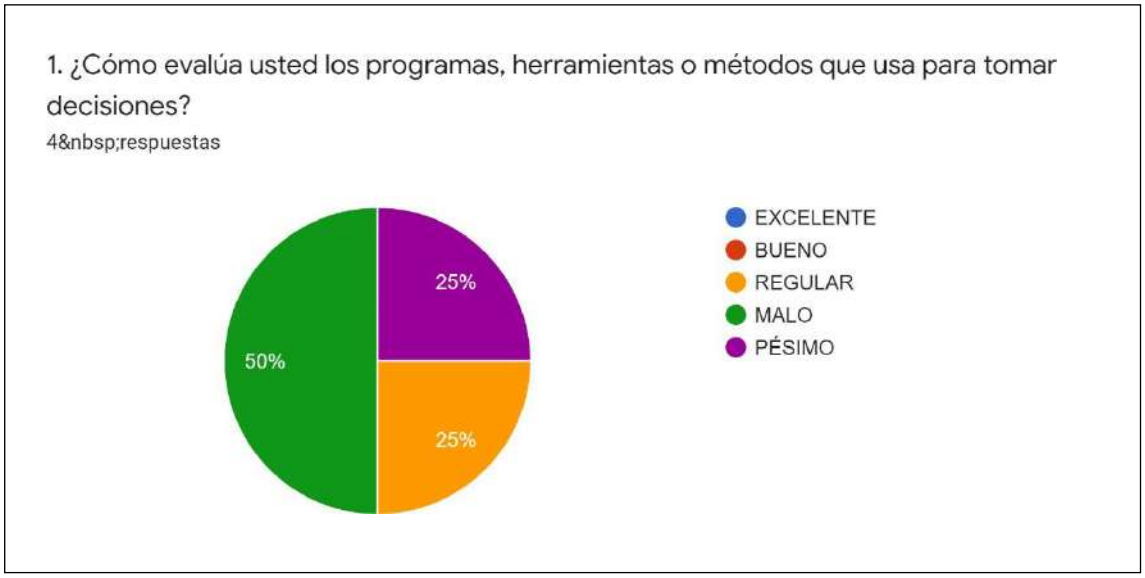
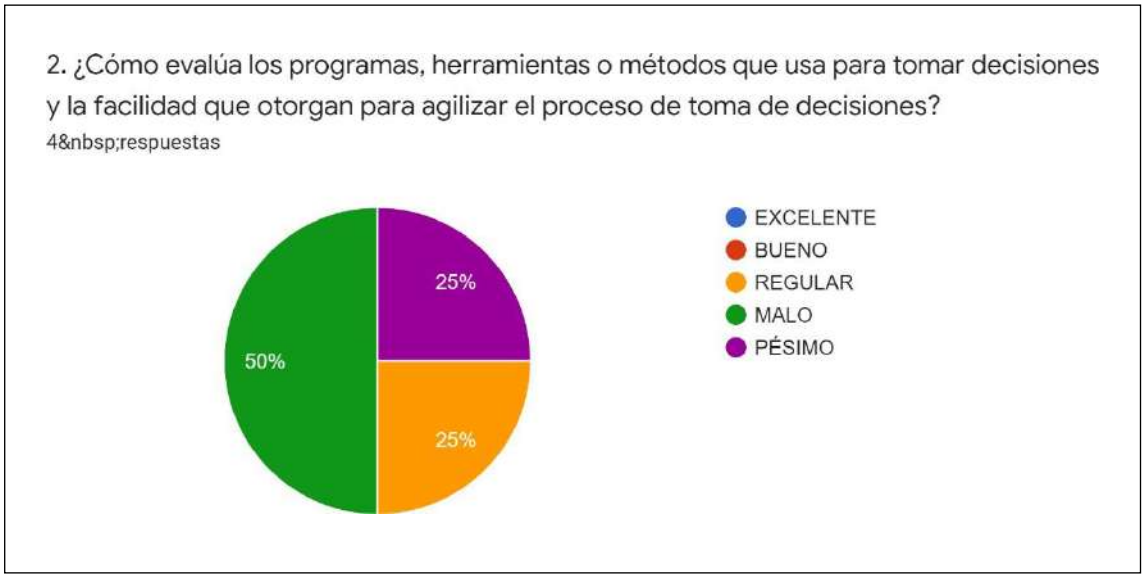
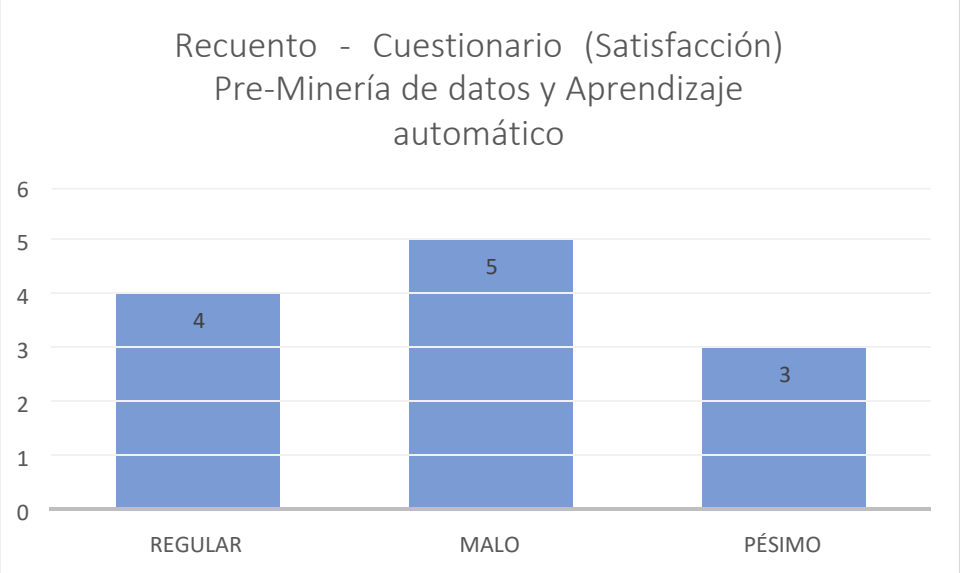

Recuento - Cuestionario (Eficiencia) Pre-Minería de datos y Aprendizaje automático
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
2
4
2
MUY POCO
POCO
REGULAR

Imagen 133: Resultados de las encuestas pre-minería de datos y aprendizaje automático – Satisfacción



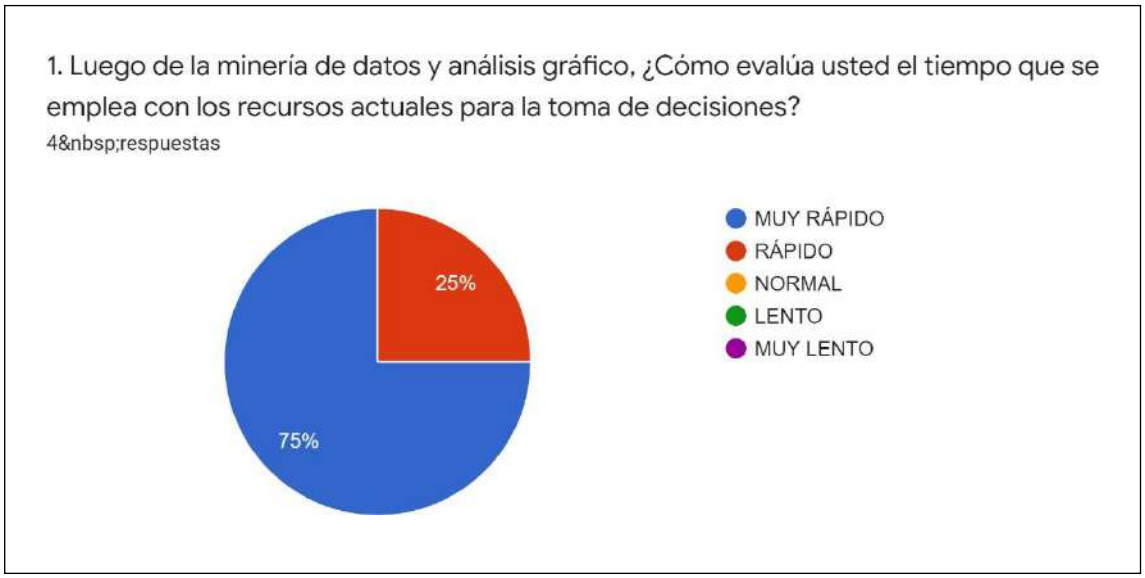


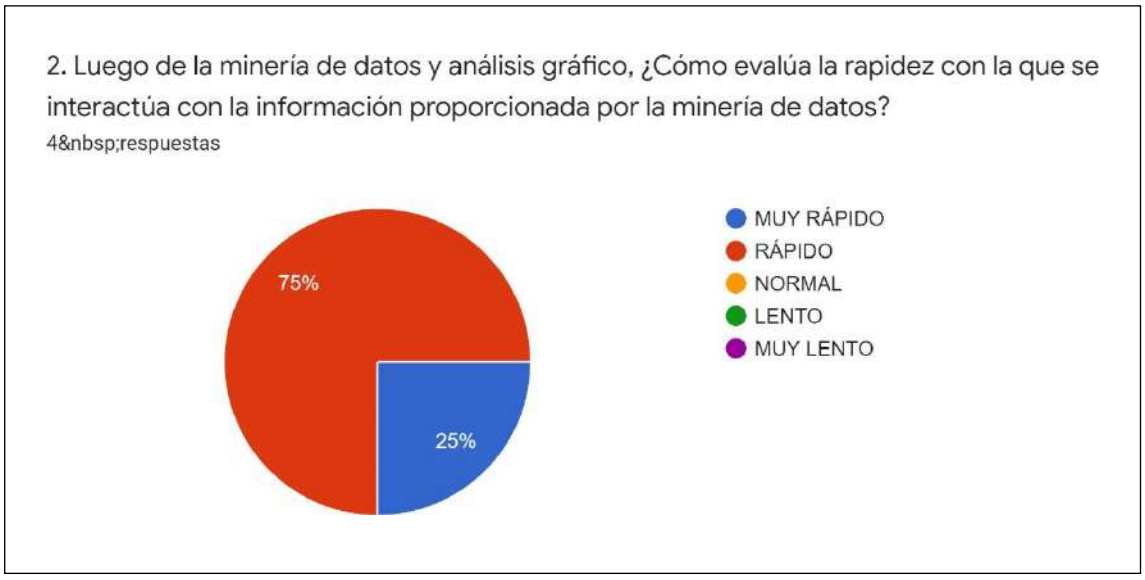

3. ¿Cómo evalúa a los programas, herramientas o métodos que usa para la toma de decisiones y la capacidad que tienen para satisfacer los requerimientos de su trabajo?
4 respuestas
25%
25%
50%
EXCELENTE
BUENO
REGULAR
MALO
PÉSIMO


Recuento - Cuestionario (Satisfacción) Pre-Minería de datos y Aprendizaje automático
6
5
4
3
2
1
0
4
5
3
REGULAR
MALO
PÉSIMO

4.3.2. RESULTADOS POST MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO

Imagen 134: Resultados de las encuestas post minería de datos y aprendizaje automático - Eficacia





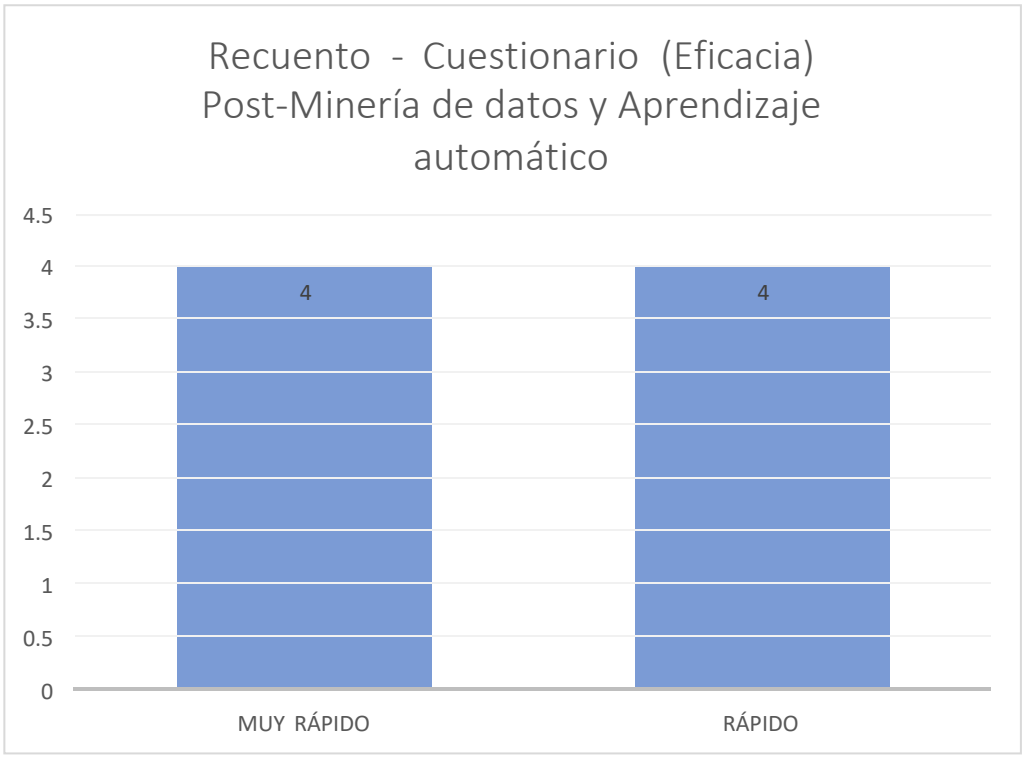

Recuento - Cuestionario (Eficacia) Post-Minería de datos y Aprendizaje automático
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
4
4
MUY RÁPIDO
RÁPIDO

Imagen 135: Resultados de las encuestas post minería de datos y aprendizaje automático - Eficiencia

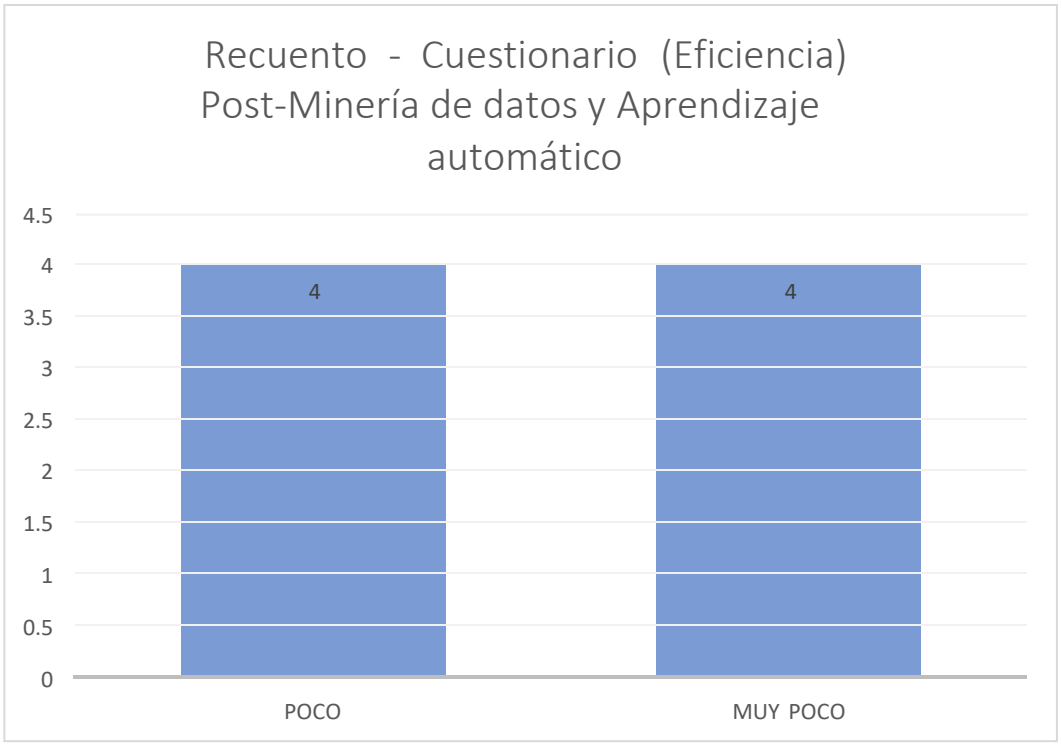
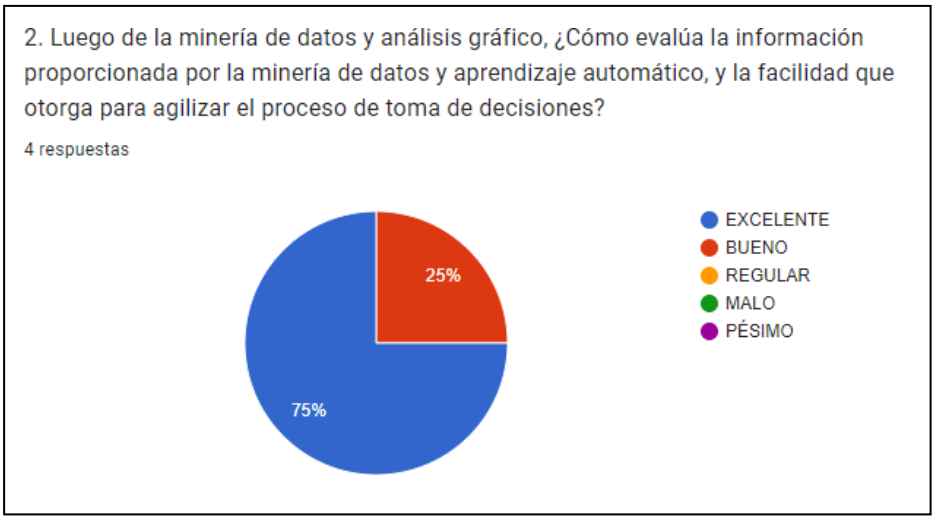
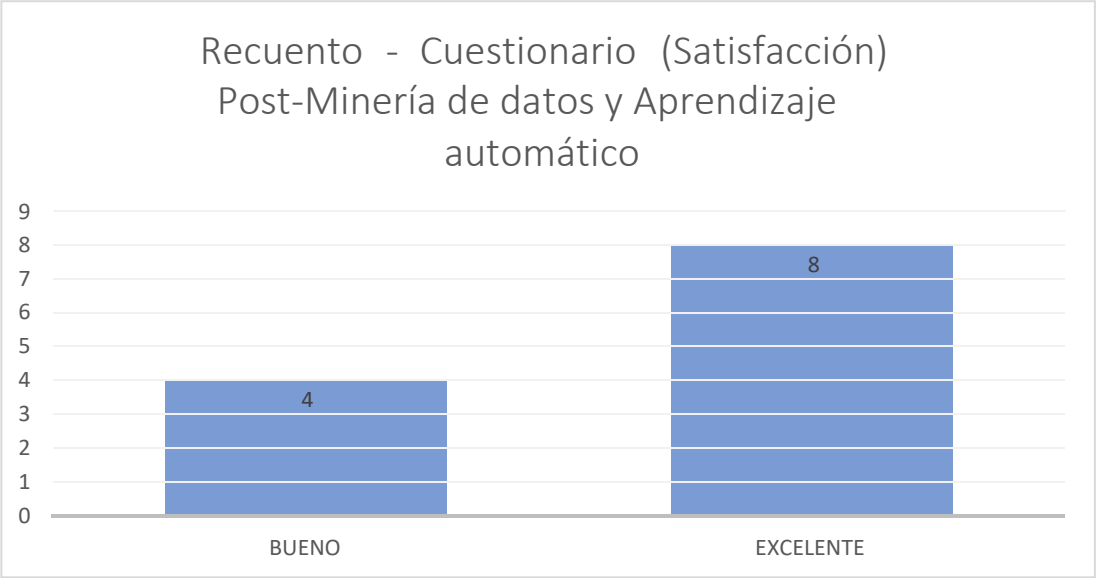

Recuento - Cuestionario (Eficiencia) Post-Minería de datos y Aprendizaje automático
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
4
4
POCO
MUY POCO

Imagen 136: Resultados de las encuestas post minería de datos y aprendizaje automático – Satisfacción





A través de los gráficos mostrados, se pude denotar el incremento en cuanto a la satisfacción de los tomadores de decisiones frente al proceso de la toma de decisiones apoyada en la minería de datos y aprendizaje automático, factor que no se apreciaba antes en los expertos antes de la presente investigación. A continuación, haremos uso de la distribución t de Student para la validación estadística de la hipótesis:

Tabla 32: Valores pre y post minería de datos y aprendizaje automático para aplicar t-student

	PRE – MINERÍA Y APREN. AUTOMÁTICO			POST – MINERÍA Y APREN. AUTOMÁTICO	
	REGULAR	MALO	PÉSIMO	RÁPIDO / POCO / BUENO	MUY RÁPIDO / MUY POCO / EXCELENTE
P1	3	2	3	4	4
P2	2	4	2	4	4
P3	4	5	3	4	8

Tabla 33: Prueba t-student (P1)

P1:	***Variable 1***	***Variable 2***
Media	2.66666667	4
Varianza	0.33333333	0
Observaciones	3	2
Varianza agrupada	0.22222222	
Diferencia hipotética de las medias	0	
Grados de libertad	3	
Estadístico t	-3.09838668	

Tabla 34: Prueba t-student (P2)

P2:	*Variable 1*	*Variable 2*
Media	2.66666667	4
Varianza	1.33333333	0
Observaciones	3	2
Varianza agrupada	0.88888889	
Diferencia hipotética de las medias	0	
Grados de libertad	3	
Estadístico t	-1.54919334	

Tabla 35: Prueba t-student (P3)

P3:	*Variable 1*	*Variable 2*
Media	4	6
Varianza	1	8
Observaciones	3	2
Varianza agrupada	3.33333333	
Diferencia hipotética de las medias	0	
Grados de libertad	3	
Estadístico t	-1.2	

Comprobamos gracias a la aplicación de la prueba t-student, que se rechaza la hipótesis nula (H0), por tanto, **se acepta la hipótesis alternativa (H1).**

CAPÍTULO V:
DISCUSIÓN DE LOS RESULTADOS

CAPÍTULO V: DISCUSIÓN DE LOS RESULTADOS

5.1. DISCUSIÓN 1

Antecedente 1

Aplicación de la minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales.

Autores: Miguel Grández.

Análisis:

El trabajo del autor señalado anteriormente tiene como objetivo analizar los datos de una distribuidora de suplementos nutricionales en la región de Lima y determinar las tendencias de compra para de esta forma formar paquetes, establecer otra estrategia de marketing y determinar qué técnica resulta más conveniente para la predicción. Toman distintas variables desde la estatura, la actividad física hasta la cantidad de hijos o el ingreso mensual; los datos iniciales los obtienen de un archivo plano.

En comparación a la obtención y depuración de los datos iniciales, el presente trabajo realiza una conversión de tablas dinámicas de un archivo en Excel a un archivo en plano para trabajar con distintos datasets según sea la conveniencia en el uso de variables, mientras que él usa un archivo plano desde el inicio y no tiene la necesidad de una depuración tan exhaustiva como la realizada en este trabajo.

5.2. DISCUSIÓN 2

Antecedente 2

Clasificación de cultivos de quinua orgánica mediante el uso de imágenes aéreas multiespectrales y técnicas de aprendizaje automática.

Autores: Donato Flores.

Análisis:

En el trabajo del autor dado a conocer anteriormente, trata sobre el uso de distintas técnicas de clasificación como son árboles de decisión, análisis discriminante, máquinas de vectores de soporte, K-means, clasificadores de conjuntos y métodos de aprendizaje profundos. Al final del trabajo se puede determinar que los modelos son efectivos y tienen la capacidad de predecir los rendimientos y, por tanto, ayudar al mapeo de cultivos de quinua.

En comparación al uso de árbol de decisión y su precisión, el presente trabajo una partición del 70% para el entrenamiento y de un 30% para la prueba por lo que nuestra cantidad de datos original que suma más de 17,000 datos, se reduce a un aproximado de 500 registros por lo no solo la precisión del árbol se ve reducida hasta un 80,9% y una coherencia de los datos del 59,6% en comparación del trabajo presentado en el antecedente donde el árbol final presenta una precisión del 91,2%; aunque en ambos casos la precisión llega a ser aceptada y el modelo se considera fiable para predecir.

5.3. DISCUSIÓN 3

Antecedente 3

Reglas de asociación y predicción utilizando series de tiempo.

Autores: Eduardo Araujo.

Análisis:

El trabajo del autor anteriormente mencionado hace uso de reglas de asociación para apoyar su objetivo principal que es la ayuda a pequeños agricultores de cultivos como choclo, palmito, lechuga y fréjol, de esta manera usa series de tiempo para la predicción. Para las reglas de asociación hace uso del algoritmo “A priori” para cada uno de los cultivos mencionados presenta tablas que reflejan la diferencia entre la confianza de un sistema usado por el Ministerio de Agricultura de su país con el programa Weka

En comparación con el software usado y la técnica empleado, el presente trabajo usa la técnica “FP-Growth” haciendo uso del programa Orange, mismo que emplea este algoritmo para su nodo de reglas de asociación; a diferencia del trabajo mencionado, la confianza varía mucho, debido a que esta investigación no presenta confianzas tan altas, como el trabajo del autor, esto debido a la cantidad de datos, ya que, nosotros usamos más de 17,000 datos en contraste de los 3682 que se usa en el mencionado trabajo. El trabajo actual presenta una confianza mínima del 84.6% y baja hasta la última con 72.3%, presentando una media de confianza de 76,9%.

5.4. DISCUSIÓN 4

Antecedente 4

Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka negra y la productividad en cultivos agrícolas.

Autores: L. Calvo-Valverde, S. Arguello, J. Guzmán-Álvarez, M. Guzmán-Quesada, M. Gonzáles-Zúñiga.

Análisis:

El trabajo de los autores citados anteriormente tiene como objetivo demostrar que las redes bayesianas dinámicas son capaces o no de predecir el avance de la plaga denominada Sigatoka negra en los cultivos de banano, por tanto, aplican el método utilizando la data histórica del clima, sin embargo, mantienen dos grupos por separado y los visualizan de manera micro y macro.

En comparación a la técnica usada y la precisión obtenida, el presente trabajo usa redes bayesianas usando árbol TAN y con un nivel de significancia del 1% obteniendo en el análisis final una precisión del modelo del 85,74%, en contraste con el trabajo de los autores mencionados que usan redes bayesianas dinámicas (RBDs) y hacen uso de un nivel de significancia del 5% obteniendo a nivel micro una precisión del 49% y a nivel macro de un 16%. Por lo que, a pesar de que son trabajos con una data distinta y enfocada a otro tipo de cultivo, el método se mantiene en un contexto de problemática en el ámbito agrícola.

5.5. DISCUSIÓN 5

Antecedente 5

Aplicación de técnicas de minería de datos para pronósticos del sector agrícola.

Autores: Tania Zamora.

Análisis:

El trabajo del autor mencionado previamente tiene como objetivo probar distintas técnicas de minería de datos como son regresión, series de tiempo, SVM y redes neuronales; justamente este último modelo es el que más resulta capaz frente a los demás y con diferencia, aunque al final del trabajo la conclusión fuese que ninguno de las técnicas usadas es suficientemente eficaz para pronosticar la producción ni la superficie en donde siembre los cultivos agrícolas como lo son la papa y el trigo de zonas determinadas que se detallan en la investigación.

A diferencia del presente trabajo, ella presenta resultados gráficos, que, si bien son perfectamente interpretables, pero solo lo hace de la red neuronal a pesar de haber probado otras tres técnicas de minería de datos; nuestro trabajo hace uso de gráficos para cada una de las técnicas que se emplearon para la resolución de los problemas que tiene la empresa Agroberries S.A.C. con el objetivo de que el tomador final de decisiones se sienta cómodo y pueda apoyarse en estos gráficos resultantes de un proceso de minería de datos, para así realizar su labor de una manera más eficaz.

CONCLUSIONES

1. Se realizó un análisis de los datos proporcionados por la empresa agroindustrial Agroberries S.A.C. En primera instancia nos encontramos con un archivo en Excel con tablas dinámicas que no podía ser transformado a un archivo plano, por tanto, se tuvo que realizar una limpieza y posterior depuración con las herramientas de Power BI y DAX Studio, de esta manera pudimos empezar a formar los datasets en archivos CSV con aquellas variables que serían determinantes para absolver la problemática que se presenta en las áreas de la empresa previamente mencionadas.

2. Se utilizó la técnica de árbol de decisión como modelo para la predicción del sobrecosto usando ciertas variables que eran las más importantes a considerar, siendo esta técnica de aprendizaje supervisado que usa la regresión y a su vez ayudando a predecir el valor continuo; se usó Python en Google Colab. La elección final del software fue Knime, que se impuso sobre Weka por la variedad altísima de herramientas que proporcionaba visual y técnicamente; de esta forma, usando el 70% de los datos para entrenamiento y el 30% de los datos para prueba, se obtuvo una ***precisión*** del 80,9% en el árbol con una ***coherencia de los datos (Kappa)*** del 59,6% y se considera fiable para los intereses de la empresa y la información que se puede obtener con respecto a las variables elegidas en primera instancia. Del mismo modo, se hallaron varios patrones que predicen con alta tasa de probabilidad los eventos de sobrecostos con respecto a las características tomadas.

3. Se hizo la implementación de la técnica de preprocesamiento Binning y se usaron reglas de asociación como la técnica de minería de datos para hallar relaciones entre los valores agrupados y el cumplimiento de Kias por cosecha, siendo esta técnica de aprendizaje no supervisado que a su vez ayuda a encontrar patrones entre las relaciones de datos, nuevamente usando Python sobre Google Colab"; finalmente se usa el software Weka en el cual se obtuvieron reglas de asociación no favorables para la empresa, es entonces que se opta por usar el software Orange que usa el algoritmo "FP-Growth" en su nodo denominado "Association Rules" (reglas de asociación) a diferencia de Weka en el que usó "A priori". Para la configuración final y la obtención de las mejores reglas de asociación se obtuvo una ***confianza*** mínima un 70% y ***un soporte mínimo*** del 0,05. De igual manera que se hallaron las reglas de asociación que predicen el cumplimiento de Kias en escenarios con características en particular, detallando el ***lift*** para cada una de estas reglas, varias de estas con una dependencia alta para las variables elegidas llegando incluso por encima del 200% lo que indica que se puede confiar en la técnica aplicada para cada uno de estos patrones encontrados.

4. Se usó el modelo de redes bayesianas para la predicción del cumplimiento de los objetivos trazados con respecto a la cantidad de arándanos esperados por cosecha; siento este modelo una técnica de aprendizaje supervisado usando la regresión para hallar el valor continuo que tomamos en cuenta en este objetivo. Luego, se usó el software SPPS Modeler 18, se procedió a usar el nodo de red bayesiana configurándolo para que haga uso del ***árbol TAN*** con un ***nivel de significancia*** del 1%, obteniendo finalmente una ***precisión del modelo*** de 85,74%, lo cual resulta ser un modelo fiable para los objetivos de la empresa. Finalmente, también se hallaron patrones que involucran a variables como fundos, valoración de la cosecha, mano de obra, jornada laboral, entre otras para hallar si se cumplirá o no el objetivo de la cosecha en particular para un determinado tiempo, mismos que resuelven la falta de información relevante sobre reportes diarios a fin de predecir con una alta probabilidad si el objetivo trazado será cubierto o no.

5. Se realizó los gráficos solicitados por los tomadores de decisiones para evaluar que tan efectiva puede ser una minería de datos en el proceso de toma de decisiones en las áreas de producción y transporte de la empresa Agroberries S.A.C., para el árbol de decisión se decidió presentarlo en informes usando la tabulación cruzada (apoyada en el porcentaje de prueba), para las reglas de asociación se presentó en modo de gráficos en barras donde se muestre claramente las mejores reglas con su respectiva confianza en cada una de ellas y, por último, para las redes bayesianas se optó por usar mapas de calor.

RECOMENDACIONES

1. Se recomienda la exploración de otras técnicas de minería de datos y aprendizaje automático para poder evaluar los resultados obtenidos en contraste con los resultados encontrados en el presente trabajo.

2. Se recomienda realizar una interfaz que permita la selección de estas y otras técnicas de minería de datos y aprendizaje automático con la finalidad de que sea mucho más amigable, intuitiva y fácil de usar para los tomadores de decisiones.

3. Se recomienda promover la aplicación del uso de técnicas de minería de datos y aprendizaje automático en otras áreas de la empresa como el área de comercialización (ventas) y que no solamente vean cultivos de arándanos, sino que se expanda de tal forma que el proceso de toma de decisiones sea más sustentado y rápido.

4. Se recomienda hacer uso de técnicas de aprendizaje profundo con fin de evaluar qué resultados se obtienen y poder ampliar el panorama de oportunidades que se nos ofrece.

REFERENCIAS BIBLIOGRÁFICAS

Acevedo, N. (2020). *Árboles de decisiones. La matemática y estadística detrás de ellos*. https://nataliaacevedo.com/arboles-de-decisiones-la-matematica-y-estadistica-detras-de-ellos/

Aekwarangkoon, S., & Thanathamathee, P. (2022). Associated Patterns and Predicting Model of Life Trauma, Depression, and Suicide Using Ensemble Machine Learning. *Emerging Science Journal*, *6*(4), 679–693. https://doi.org/10.28991/esj-2022-06-04-02

Agromedo Cueva, G., & Salazar Ávila, E. (2019). *Inteligencia de negocios para la agilización en la toma de decisiones de la empresa industrial CAMPOSOL S.A.* [UNT]. https://dspace.unitru.edu.pe/bitstream/handle/UNITRU/12670/Argomedo Cueva, Gemma Yaquelyn; Salazar Ávila, Erika Isabel.pdf?sequence=1&isAllowed=y

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-0177-7

Araujo, E. (2018). Reglas De Asociación Y Predicción Utilizando Series De Tiempo. In *Tesis* (Vol. 1). http://dspace.ups.edu.ec/bitstream/123456789/5081/1/UPS-CYT00109.pdf

Ayele, W. Y. (2020). Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Science and Applications*, *11*(6), 20–32. https://doi.org/10.14569/IJACSA.2020.0110603

Báez, J., Paredes, C., Sosa, G., & García, M. (2018). *Descubriendo reglas de asociaci´on en bases de datos del sector retail usando R*. http://sedici.unlp.edu.ar/bitstream/handle/10915/73220/Documento_completo.pdf?sequence=1

Bodero-Poveda, E. I., Morales-Alarcón, C. I., & Estefanía Ramos-Araujo III, C. (2022). Ciencias Técnicas y Aplicadas Artículo de Investigación Técnicas de minería de datos para el análisis de la plusvalía inmobiliaria Data mining techniques for real estate appreciation analysis Técnicas de mineração de dados para análise de ganhos de capita. *Núm. 1. Enero-Marzo*, *8*, 916–930. http://dominiodelasciencias.com/ojs/index.php/es/index

Calvo-Valverde, L. A., Argüello, S., Guzmán-Alvarez, J. A., Guzmán-Quesada, M., & González-Zúñiga, M. (2019). Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka negra y la productividad en cultivos agrícolas. *Revista Tecnología En Marcha*, *32*, 158–170. https://doi.org/10.18845/tm.v32i4.4800

Cazacu, M., & Titan, E. (2020). Adapting CRISP-DM for Social Sciences. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *11*(2sup1), 99–106. https://doi.org/10.18662/brain/11.2sup1/97

Ccoyccosi, R., Huanay, L., Huayllasco, E., Loayza, V., & Mayorga, K. (2021). *Propuesta de un modelo de machine learning para el pronóstico de la demanda de prendas de vestir en la Corporación Brusko S.A.C.* 1–84. https://repositorio.esan.edu.pe/bitstream/handle/20.500.12640/2931/2021_IIC_21-2_04_TC.pdf?sequence=1&isAllowed=y

Chandra, P. (2020). *Partitioning Method (K-Mean) in Data Mining*. https://www.geeksforgeeks.org/partitioning-method-k-mean-in-data-mining/

Chhaya, K., Khanzode, A., & Sarode, R. D. (2020). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literate Review. *International Journal of Library & Information Science (IJLIS)*, *9*, 3. http://www.iaeme.com/IJLIS/index.asp30http://www.iaeme.com/IJLIS/issues.asp?JType=IJLIS&VType=9&IType=1JournalImpactFactor%0Awww.jifactor.comhttp://www.iaeme.com/IJLIS/issues.asp?JType=IJLIS&VType=9&IType=1%0Ahttp://iaeme.com

Del Río Cárdenas, S. A. (2019). *Inferencia de Interacciones Causales Génicas Usando Técnicas basadas en Manto de Markov*. https://core.ac.uk/download/pdf/250405346.pdf

Ebrahim, M. (2019). *Tutorial de Python pandas: Iniciando con DataFrames*. https://likegeeks.com/es/tutorial-de-python-pandas/

Edastama, P., Bist, A. S., & Prambudi, A. (2021). Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm. *International Journal of Cyber and IT Service Management*, *1*(2), 159–172. https://doi.org/10.34306/ijcitsm.v1i2.46

Fátima, D., & Serrano, O. (n.d.). *TÉCNICAS ESTADÍSTICAS APLICADAS EN NUTRICIÓN Y SALUD*.

Flores, D. (2021). *Clasificación de cultivos de quinua orgánica mediante el uso de imágenes aéreas multiespectrales y técnicas de aprendizaje automático*. https://tesis.pucp.edu.pe/repositorio/bitstream/handle/20.500.12404/20855/FLORES_ESPINOZA_DONATO_ANDRES_CLASIFICACION_CULTIVOS_QUINUA.pdf?sequence=1&isAllowed=y

Giustin, F. N., Sari, B. N., Padilah, T. N., Studi, P., Informatika, T., Karawang, U. S., & Author, C. (2022). *APPLICATION OF C5 . 0 ALGORITHM IN PREDICTION OF LEARNING OUTCOMES IN CALCULUS SUBJECT*. *3*(2), 90–97.

Gonzalez, L. (2021). *Algoritmo Apriori - Teoría*. https://aprendeia.com/algoritmo-apriori/

GRÁNDEZ, M. (2017). *Aplicación De Minería De Datos Para Determinar Patrones De Consumo Futuro En Clientes De Una Distribuidora De Suplementos Nutricionales* [USIL]. http://repositorio.usil.edu.pe/bitstream/USIL/2763/1/2017_Granda_Aplicacion-de-mineria-datos.pdf

Gupta, A. (2021). *Algoritmo ECLAT*. https://es.acervolima.com/2021/02/09/ml-algoritmo-eclat/

Haro, S. (2017). *TÉCNICAS DE CLASIFICACIÓN EN MINERÍA DE DATOS Y SOFTWARE ORANGE CANVAS. APLICACIÓN CON DATOS METEOROLÓGICOS*. https://masteres.ugr.es/moea/pages/curso201617/tfm1617/tfm_haroriverasilvia/!

Knuth, K. H. (2019). Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing: A Review Journal*, *95*, 102581. https://doi.org/10.1016/j.dsp.2019.102581

Leslie, Y., & Salazar, P. (2022). *Modelo de aprendizaje automático para identificar operaciones inusuales de lavado de activos en una entidad financiera*. http://repositorio.unap.edu.pe/handle/UNAP/18340

Mamgain, A. (2018). Guidance to Data Mining in Python. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT* |, *3*(10), 2456–3307.

Mardel, T. (2019). *Application of datamining techniques to the demography area of the national office of statistics and information*. http://eventos.upr.edu.cu/index.php/MARDELTUR2019/TITGD/paper/viewFile/1780/1335

Martinez, M., Escobar, B., García-Díaz, M. E., & Pinto-Roa, D. P. (2021). Market basket analysis with association rules in the retail sector using Orange. Case study: Appliances sales company. *CLEI Eletronic Journal (CLEIej)*, *24*(2), 1–17. https://doi.org/10.19153/cleiej.24.2.12

Mateos, C. (2021). *Definición y estudio de redes bayesianas aplicadas a Ciencias de la Salud y de la Vida*. https://eprints.ucm.es/id/eprint/68123/1/mateos_marcos_tfm_bioestadistica.pdf

Mesa, L., Rivera, M., & Romero, J. (2018). *Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión*. https://www.urosario.edu.co/Administracion/documentos/investigacion/laboratorio/miller_2_2.pdf

Minitab, L. (2019). *Introduction to Data Binning*. www.minitab.com.

Molina, J., & García, J. (2006). *TÉCNICAS DE ANÁLISIS DE DATOS*. http://matema.ujaen.es/jnavas/web_recursos/archivos/weka

Monteserin, A. (2018). *Reglas de asociación*. https://users.exa.unicen.edu.ar/catedras/optia/public_html/2018

Muñoz, C. G. (2022). *Aplicación de técnicas binning en geoestadística no paramétrica*. 1–73.

Navas, F. (2016). *INTRODUCCIÓN A LA MINERÍA DE DATOS CON WEKA: APLICACIÓN A UN PROBLEMA ECONÓMICO*. http://tauja.ujaen.es/bitstream/10953.1/6984/1/TFG

Papathanasiou, J., Belioka, M. P., Dikoglou, P., & Zopounidis, D. (2022). Proceedings of the 8th International Conference on Composite Structures. In *Composite Structures* (2022nd ed., Vol. 32, Issues 1–4, pp. 200–200).

https://books.google.es/books?hl=es&lr=&id=7Kl3EAAAQBAJ&oi=fnd&pg=PA200&dq=KDD+steps&ots=ozBau1sfmN&sig=OM25JdQT3KgdwfM_B7cA7MSZAQc#v=onepage&q&f=false

Pérez, C. (2018). *REGLAS DE ASOCIACIÓN. APLICACIÓN PRÁCTICA EN LA CESTA DE LA COMPRA DE LOS CONSUMIDORES.* https://idus.us.es/bitstream/handle/11441/88401/Reglas_de_asociacion.pdf?sequence=1&isAllowed=y

Plotnikova, V., Dumas, M., Nolte, A., & Milani, F. (2022). Designing a data mining process for the financial services domain ABSTRACT. *Journal of Business Analytics*, *00*(00), 1–27. https://doi.org/10.1080/2573234X.2022.2088412

Quintana, M. (2017). *Capítulo 6: Desviación Estandar*. https://docplayer.es/66936128-Universidad-nacional-del-callao-facultad-de-ciencias-administrativas.html

Ramírez, P. (2020). *Redes Bayesianas para predicción y descubrimiento de relaciones con señales procedentes de sensores industriales*. https://repositorio.uam.es/bitstream/handle/10486/692549/ramirez_hereza_pablo_tfm.pdf?sequence=1

Romero, B. (2020). *Una introducción a los modelos de Machine Learning*. https://repositorioinstitucional.buap.mx/handle/20.500.12371/10527

Ruiz, L. (2020). *Medidas de Tendencia Central*. https://www.uaeh.edu.mx/division_academica/educacion-media/repositorio/2010/6-semestre/estadistica/medidas-tendencia-central.pdf

Sáenz, A., Cortés, F., & Betancourt, J. (2017). *Reglas de asociación en una Base de datos del área médica*. https://www.redalyc.org/pdf/1939/193954081005.pdf

Sarker, I. H. (2021). Machine Learning: Algorithms , Real - World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Soler-Flores, F., González-Cancela, N., & Serrano, B. M. (2019). *Aplicación de redes bayesianas para el control de la frecuencia de los accidentes viarios Application of Bayesian Networks to Control Road Accident the Frequency*.

Soler, M. (2017). *Reglas de Asociación en Weka*. https://docplayer.es/43555598-Reglas-de-asociacion-en-weka.html

Soto, J., & Martínez, M. (2020). *Aprendizaje Supervisado: Regresión y Clasificación en KNIME*. https://abierta.ugr.es/pluginfile.php/27691/mod_resource/content/1/8.3.pdf

Suarez, A. J. B., Singh, B., Almukhtar, F. H., Kler, R., Vyas, S., & Kaliyaperumal, K. (2022). Identifying Smart Strategies for Effective Agriculture Solution Using Data Mining Techniques. *Journal of Food Quality*, *2022*. https://doi.org/10.1155/2022/6600049

Tapia, I. (2015). *Manual Básico Knime*. https://cupdf.com/document/manual-basico-knime.html

Urbina-Nájera, A. B., & Méndez-Ortega, L. A. (2022). Predictive Model for Taking Decision to Prevent University Dropout. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(4), 205–213. https://doi.org/10.9781/ijimai.2022.01.006

Wang, J. (2022). Mining and Prediction of Large Sport Tournament Data Based on Bayesian Network Models for Online Data. *Wireless Communications and Mobile Computing*, *2022*. https://doi.org/10.1155/2022/1211015

Yash V. Bagal. (2020). Data Mining in Agriculture-A Novel Approach. *International Journal of Engineering Research And*, *V9*(08), 213–215. https://doi.org/10.17577/ijertv9is080107

Zamora, T. (2018). *APLICACIÓN DE TECNICAS DE MINERIA DE DATOS PARA PRONÓSTICOS DEL SECTOR AGRÍCOLA*. http://opac.pucv.cl/pucv_txt/txt-8000/UCC8100_01.pdf

ANEXOS

ANEXO 1. Encuesta pre-minería de datos y aprendizaje automático.

Cuestionario Pre-Minería de datos y Aprendizaje automático

A continuación, se presentan preguntas acerca de la satisfacción que se tiene con respecto al proceso de toma de decisiones y los métodos, herramientas y/o programas que interviene para llevar a cabo este proceso.

EFICIENCIA

1. ¿Cómo evalúa usted el tiempo que se emplea con los recursos actuales para la toma de decisiones?

Marca solo un óvalo.

- MUY RÁPIDO
- RÁPIDO
- NORMAL
- LENTO
- MUY LENTO

2. ¿Cómo evalúa la rapidez con la que se interactúa con la metodología, herramienta o programa que usa para las tareas del proceso de toma de decisiones?

Marca solo un óvalo.

- MUY RÁPIDO
- RÁPIDO
- NORMAL
- LENTO
- MUY LENTO

EFICACIA

1. ¿Cómo evalúa la cantidad de documentación que se tienen que usar junto a la herramienta, metodología o programa que se emplea para la toma de decisiones?

Marca solo un óvalo.

- DEMASIADO
- CONSIDERABLE
- REGULAR
- POCO
- MUY POCO

2. ¿Cómo evalúa la cantidad de tecnologías de información que se usan para el proceso de toma de decisiones?

Marca solo un óvalo.

- DEMASIADO
- CONSIDERABLE
- REGULAR
- POCO
- MUY POCO

SATISFACCIÓN

1. ¿Cómo evalúa usted los programas, herramientas o métodos que usa para tomar decisiones?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
-

PÉSIMO

2. ¿Cómo evalúa los programas, herramientas o métodos que usa para tomar decisiones y la facilidad que otorgan para agilizar el proceso de toma de decisiones?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
- PÉSIMO

3. ¿Cómo evalúa a los programas, herramientas o métodos que usa para la toma de decisiones y la capacidad que tienen para satisfacer los requerimientos de su trabajo?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
- PÉSIMO

ANEXO 2. Encuesta post minería de datos y aprendizaje automático.

Cuestionario Post-Minería de datos

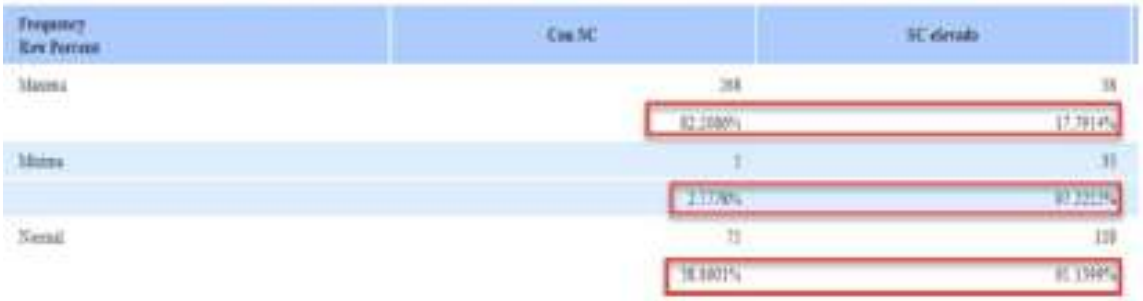
A continuación, se presentan preguntas relacionadas a los resultados en el periodo de cosecha de arándanos en la empresa Agroberries S.A.C. en el año 2020 y que corresponden a la cosecha diaria, kias y cumplimiento de los objetivos trazados.

División de secciones

En el siguiente cuestionario existen 4 secciones, los cuales se dividen de la siguiente manera:
- SECCIÓN 1: Con respecto al sobrecosto que existe en las cosechas de arándanos.
- SECCIÓN 2: Con respecto al cumplimiento de Kias.
- SECCIÓN 3: Con respecto al objetivo trazado de la cantidad de arándanos cosechados.
- SECCIÓN 4: Preguntas finales.

SECCIÓN 1: ÁRBOL DE DECISIÓN	Presentación de resultados y preguntas acerca de la efectividad / eficiencia que da la minería de datos con respecto a la toma de decisiones y evitar SOBRECOSTO EN LA COSECHA DE ARÁNDANOS.

JORNADA LABORAL con respecto al SOBRECOSTO:



- Existe un 17,79% de posibilidad que exista SOBRECOSTO ELEVADO en la cosecha si la JORNADA LABORAL es MÁXIMA.
- Existe un 82,21% de posibilidad que existe SOBRECOSTO (NO ELEVADO) en la cosecha si la JORNADA LABORAL es MÁXIMA.

- Existe un 97,22% de posibilidad que exista SOBRECOSTO ELEVADO en la cosecha si la JORNADA LABORAL es MÍNIMA.
- Existe un 2,788% de posibilidad que existe SOBRECOSTO (NO ELEVADO) en la cosecha si la JORNADA LABORAL eS MÍNIMA.

- Existe un 61,14% de posibilidad que exista SOBRECOSTO ELEVADO en la cosecha si la JORNADA LABORAL es NORMAL.
- Existe un 38,86% de posibilidad que exista SOBRECOSTO (NO ELEVADO) en la cosecha si la JORNADA LABORAL es NORMAL.

VARIEDAD DE ARÁNDANO COSECHADA con respecto al SOBRECOSTO:

Cross Tabulation of VariedadCosechada by Sobrecosto

Frequency Row Percent	Con SC	SC elevado
Atlas	44	19
	69.8413%	30.1587%
Biloxi	62	98
	38.75%	61.25%
Jupiter	3	
	100%	
Sekoya Beauty	5	6
	45.4545%	54.5455%
Sekoya Pop	18	13
	58.0645%	41.9355%
Ventura	214	79
	74.0484%	25.9516%

Se destacan las más resaltantes:

- Existe un 69,84% de posibilidades que si la VARIEDAD COSECHADA es ATLAS, exista SOBRECOSTO (NO ELEVADO).
- Existe un 61,26% de posibilidades que si la VARIEDAD COSECHADA es BILOXI, exista SOBRECOSTO ELEVADO.
- Existe un 64,66% de posibilidades que si la VARIEDAD COSECHADA es SEKOYA BEAUTY, exista SOBRECOSTO ELEVADO.
- Existe un 68,06% de posibilidades que si la VARIEDAD COSECHADA es SEKOYA POP, exista SOBRECOSTO (NO ELEVADO).
- Existe un 74,06% de posibilidades que si la VARIEDAD COSECHADA es VENTURA, exista SOBRECOSTO (NO ELEVADO).

VALORACIÓN COSECHA (cantidad cosechada) con respecto al SOBRECOSTO:

Frequency Row Percent	Con SC	SC elevado
Buena	0	50
	25.0598%	74.9402%
Mala	144	80
	61.5385%	38.4615%
Normal	281	81
	67.7108%	32.2892%

- Existe un 76,92% de posibilidades que exista SOBRECOSTO ELEVADO si la VALORACIÓN DE LA COSECHA es BUENA.
- Existe un 61,64% de posibilidades que exista SOBRECOSTO (NO ELEVADO) si la VALORACIÓN DE LA COSECHA es MALA.
- Existe un 67,73% de posibilidades que exista SOBRECOSTO (NO ELEVADO) si la VALORACIÓN DE LA COSECHA es NORMAL.

CANTIDAD DE MANO DE OBRA (usada en la cosecha) con respecto al SOBRECOSTO:

Frequency Row Percent	Con SC	SC elevado
Baja MO	133	131
	57.9649%	42.0351%
Buena MO	33	33
	37.8471%	62.1529%
Regular MO	138	47
	77.1845%	22.8155%

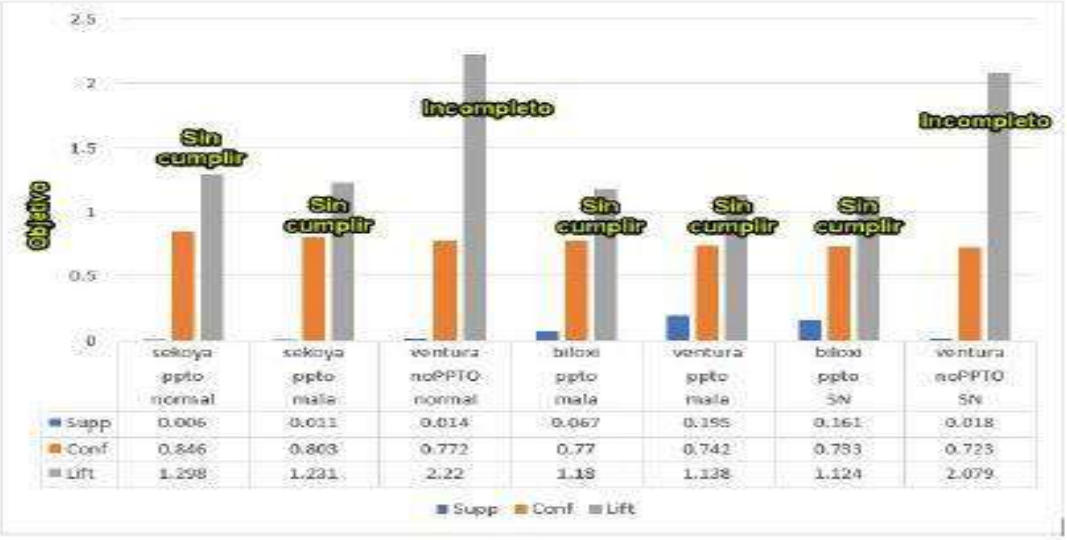
Si la MANO DE OBRA fue BAJA...
- Existe un 67,96% de posibilidades que exista SOBRECOSTO (NO ELEVADO) .
- Existe un 42,06% de posibilidades que exista SOBRECOSTO ELEVADO.

Si la MANO DE OBRA fue BUENA...
- Existe un 62,36% de posibilidades que exista SOBRECOSTO ELEVADO.
- Existe un 37,66% de posibilidades que exista SOBRECOSTO (NO ELEVADO).

Si la MANO DE OBRA fue REGULAR...
- Existe un 77,19% de posibilidades que exista SOBRECOSTO (NO ELEVADO).
- Existe un 22,82% de posibilidades que exista SOBRECOSTO ELEVADO.

SECCIÓN 2: REGLAS DE ASOCIACIÓN

Presentación de resultados y preguntas acerca de la efectividad / eficiencia que da la minería de datos con respecto a la toma de decisiones y CUMPLIMIENTO DE KIAS.



	sekoya ppto normal	sekoya ppto mala	ventura noPPTO normal	biloxi ppto mala	ventura ppto mala	biloxi ppto SN	ventura noPPTO SN
Supp	0.006	0.011	0.014	0.067	0.155	0.161	0.018
Conf	0.846	0.803	0.772	0.77	0.742	0.733	0.723
Lift	1.298	1.231	2.22	1.18	1.138	1.124	2.079

El siguiente gráfico muestra las reglas que se cumplen en la cosecha con respecto al cumplimiento de Kias y se interpretan de la siguiente manera:

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a SEKOYA, con un COSTO de cosecha PRESUPUESTADO (PPTO) y VALORACIÓN DE COSECHA (cantidad cosechada en kilos) es NORMAL, existe un 84,6% de confianza que el OBJETIVO de las KIAS para esta cosecha quede SIN CUMPLIR.

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a SEKOYA cn un COSTO de cosecha PRESUPUESTADO (PPTO) y VALORACIÓN DE COSECHA (cantidad cosechada en kilos) es MALA, existe un 80,3% de confianza que el OBJETIVO de la KIAS para esta cosecha quede SIN CUMPLIR.

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a VENTURA con un COSTO de cosecha NO PRESUPUESTADO (noPPTO) y VALORACIÓN DE COSECHA (cantidad cosechada en kilos) es NORMAL, existe un 77,2% de confianza que el OBJETIVO de las KIAS para esta cosecha quede INCOMPLETO.

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a BILOXI con un COSTO de cosecha PRESUPUESTADO (PPTO) y VALORACIÓN DE COSECHA (cantidad cosechada en kilos) es MALA, existe un 77% de confianza que el OBJETIVO de las KIAS para esta cosecha quede SIN CUMPLIR.

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a VENTURA con un COSTO de cosecha PRESUPUESTADO (PPTO) y VALORACIÓN DE COSECHA (cantidad cosechada en kilos) es MALA, existe un 74,2% de confianza que el OBJETIVO de las KIAS para esta cosecha quede SIN CUMPLIR.

Las reglas de a continuación no tienen como variable a la VALORACIÓN DE COSECHA:

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a BILOXI y con un COSTO de cosecha PRESUPUESTADO (PPTO), existe un 78,3% de confianza que el OBJETIVO de las KIAS para esta cosecha quede SIN CUMPLIR.

- Si la COSECHA tuvo como VARIEDAD DE ARÁNDANO cosechada a VENTURA y con un COSTO de cosecha NO PRESUPUESTADO (noPPTO), existe un 72,3% de confianza que el OBJETIVO de las KIAS para esta cosecha quede INCOMPLETO.

SECCIÓN 3: MAPA DE CALOR

Presentación de resultados y preguntas acerca de la efectividad / eficiencia que da la minería de datos con respecto a la toma de decisiones y LOGRAR OBJETIVOS TRAZADOS EN LA CANTIDAD DE ARÁNDANOS COSECHADOS.

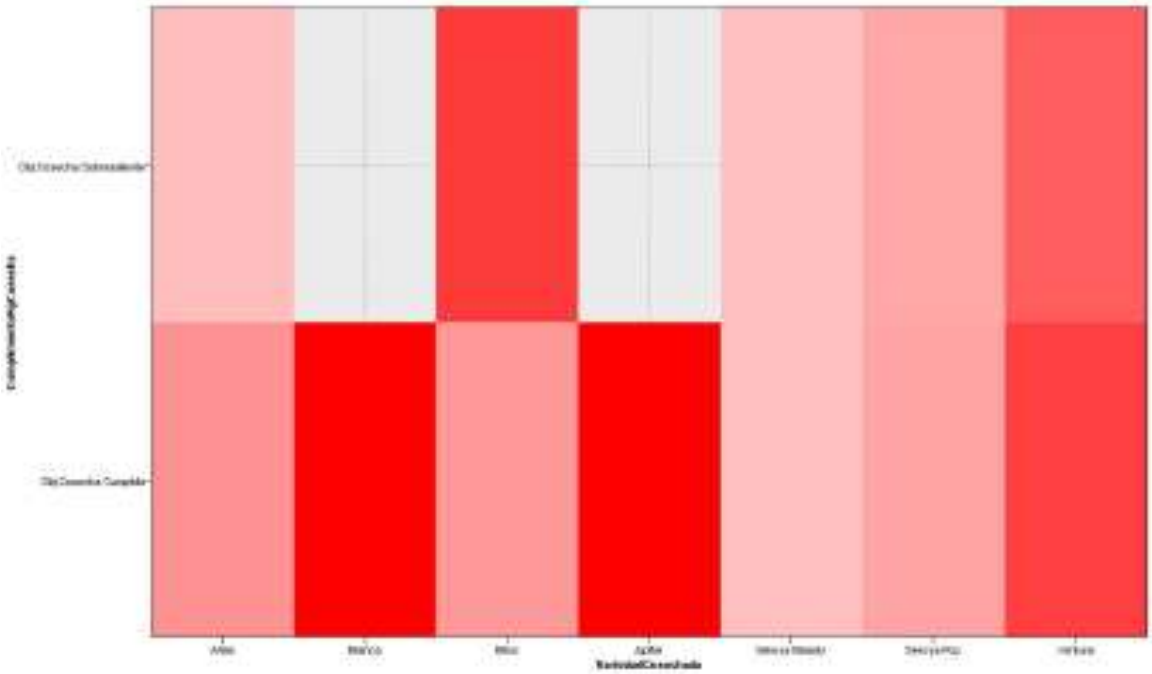
CUMPLIMIENTO DEL OBJETIVO DE KG DE COSECHA respecto a la JORNADA LABORAL

Si la JORNADA LABORAL fue MÁXIMA, existe...
- Un 89,16% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 92,76% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la JORNADA LABORAL fue NORMAL, existe...
- Un 82,61% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 76,81% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la JORNADA LABORAL fue MÍNIMA, existe...
- Un 90,27% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 78,08% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

CUMPLIMIENTO DEL OBJETIVO DE KG DE COSECHA respecto a la VARIEDAD DE ARÁNDANOS cosechados



Si la VARIEDAD DE ARÁNDANO cosechada fue BILOXI, existe...
- Un 90,76% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 76,79% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VARIEDAD DE ARÁNDANO cosechada fue ATLAS, existe...
- Un 70,23% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 77,17% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

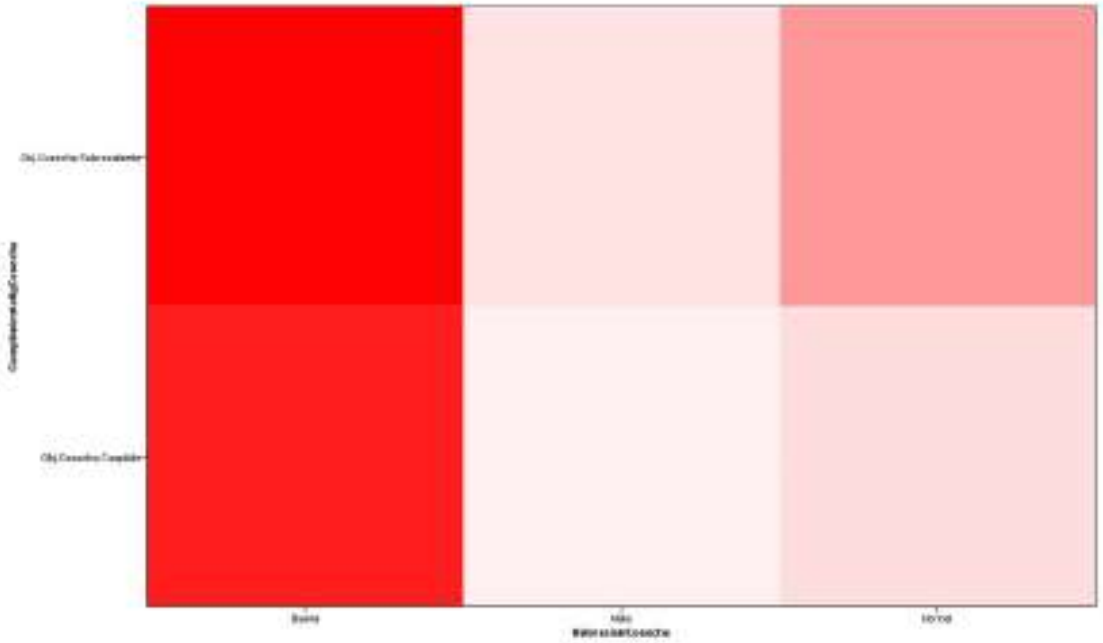
Si la VARIEDAD DE ARÁNDANO cosechada fue VENTURA, existe...
- Un 86,20% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 90,28% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VARIEDAD DE ARÁNDANO cosechada fue SEKOYA POP, existe...
- Un 73,70% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 74,11% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VARIEDAD DE ARÁNDANO cosechada fue JUPITER, existe...
- Un 100% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VARIEDAD DE ARÁNDANO cosechada fue BIANCA, existe...
- Un 100% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

CUMPLIMIENTO DEL OBJETIVO DE KG DE COSECHA respecto a la VALORACIÓN COSECHA (cantidad de kilos cosechados)

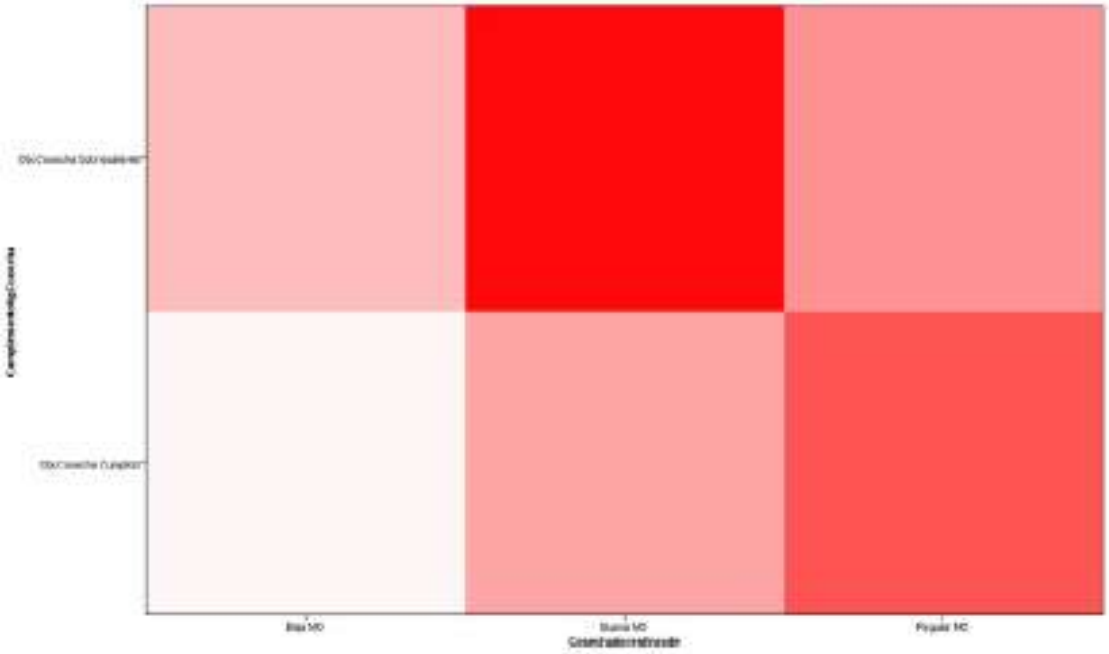


Si la VALORACIÓN DE LA COSECHA (cantidad de kilos cosechados) fue BUENA, existe...
- Un 96,76% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 94,26% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VALORACIÓN DE LA COSECHA (cantidad de kilos cosechados) fue NORMAL, existe...
- Un 87,08% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 82,88% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la VALORACIÓN DE LA COSECHA (cantidad de kilos cosechados) fue MALA, existe...
- Un 82,67% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 81,86% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

CUMPLIMIENTO DEL OBJETIVO DE KG DE COSECHA respecto a la CANTIDAD DE COSECHADORES que participaron en la cosecha



Si la CANTIDAD DE COSECHADORES que participaron en la cosecha fue BUENA M.O., existe...
- Un 93,60% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 84,91% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la CANTIDAD DE COSECHADORES que participaron en la cosecha fue REGULAR M.O., existe...
- Un 86,96% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 89,39% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

Si la CANTIDAD DE COSECHADORES que participaron en la cosecha fue BAJA M.O., existe...
- Un 83,68% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea SOBRESALIENTE.
- Un 80,46% de probabilidades que el OBJETIVO DE KGs A COSECHAR sea CUMPLIDO.

SECCIÓN 4.1: PREGUNTAS FINALES - EFICIENCIA

1. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa usted el tiempo que se emplea con los recursos actuales para la toma de decisiones?

Marca solo un óvalo.

- MUY RÁPIDO
- RÁPIDO
- NORMAL
- LENTO
- MUY LENTO

2. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa la rapidez con la que se interactúa con la información proporcionada por la minería de datos?

Marca solo un óvalo.

- MUY RÁPIDO
- RÁPIDO
- NORMAL
- LENTO
- MUY LENTO

SECCIÓN 4.2: PREGUNTAS FINALES - EFICACIA

1. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa la cantidad de documentación que se tienen que usar junto a la información proporcionada por la minería de datos?

Marca solo un óvalo.

- DEMASIADO
- CONSIDERABLE
- REGULAR
- POCO
- MUY POCO

2. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa la cantidad de tecnologías de información que se usan para el proceso de toma de decisiones?

Marca solo un óvalo.

- DEMASIADO
- CONSIDERABLE
- REGULAR
- POCO
- MUY POCO

SECCIÓN 4.3: PREGUNTAS FINALES - SATISFACCIÓN

1. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa usted la la información proporcionada por la minería de datos?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
- PÉSIMO

2. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa la información proporcionada por la minería de datos y la facilidad que otorga para agilizar el proceso de toma de decisiones?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
- PÉSIMO

3. Luego de la minería de datos y análisis gráfico, ¿Cómo evalúa la información proporcionada por la minería de datos y la capacidad que tiene para satisfacer los requerimientos de su trabajo?

Marca solo un óvalo.

- EXCELENTE
- BUENO
- REGULAR
- MALO
- PÉSIMO

ANEXO 3. Resolución del dictamen.

UPAO | Facultad de Ingeniería

Trujillo, 18 de noviembre de 2022

RESOLUCIÓN N° 2298-2022-FI-UPAO

VISTO, el informe favorable del Jurado Evaluador del Proyecto de Tesis, titulado: **"MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022"**, de los Bachilleres: **SANCHEZ OTINIANO, LUIS FERNANDO y MENDOZA VASQUEZ, JHORDYN DANIEL,** del Programa de Estudio de Ingeniería de Computación y Sistemas, y;

CONSIDERANDO:

Que, el Jurado Evaluador conformado por los señores docentes: **Ms. AGUSTIN EDUARDO ULLON RAMIREZ,** Presidente; **Ms. HEBER GERSON ABANTO CABRERA**, Secretario; **Ms. EDWARD FERNANDO CASTILLO ROBLES**, Vocal; han revisado el Proyecto de Tesis, encontrándolo conforme;

Que, el Proyecto de Tesis ha sido elaborado conforme a las exigencias prescritas por el Reglamento de Grados y Títulos de Pregrado de la Universidad, el mismo que fue sometido a evaluación por el mencionado jurado evaluador, quien por acuerdo unánime recomendó su aprobación, tal como se desprende del informe elevado a la Facultad de Ingeniería;

Que, de acuerdo al Artículo 28° del Reglamento de Grados y Títulos de la Universidad, el Proyecto de Tesis se inscribe en el libro de proyectos de tesis a cargo de la Secretaría Académica de la Facultad;

Estando al Estatuto de la Universidad, al Reglamento de Grados y Títulos la Universidad y a las atribuciones conferidas a éste Despacho;

SE RESUELVE:

PRIMERO: **APROBAR** la modalidad de titulación de los Bachilleres: **SANCHEZ OTINIANO, LUIS FERNANDO y MENDOZA VASQUEZ, JHORDYN DANIEL**, consistente en presentación, ejecución y sustentación de una **TESIS** para optar el título profesional de **INGENIERO DE COMPUTACIÓN Y SISTEMAS**.

SEGUNDO: **APROBAR y DISPONER** la inscripción del Proyecto de Tesis titulado: **"MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022".**

TERCERO: **COMUNICAR** a los Bachilleres que tienen un plazo máximo de **UN AÑO** para desarrollar su tesis, a cuyo vencimiento, se produce la caducidad del mismo, perdiendo el derecho exclusivo sobre el tema elegido.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE.

UNIVERSIDAD PRIVADA ANTENOR ORREGO FACULTAD DE INGENIERÍA DECANATO PERÚ

Dr. Ángel Alanoca Quenta
DECANO

ANEXO 4. Carta de no dilvugación de datos.

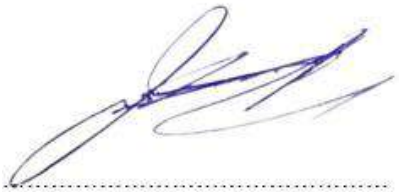
Diego Jesús Rodríguez Navarrete
Dirección: Urb. Monserrate Mz. A Lt. 12
Móvil: 952309045
diego_j21@hotmail.com
Trujillo, a 28 de septiembre de 2022

Empresa: Agroberries S.A.C.
Empleado: Diego Jesús Rodríguez Navarrete
Cargo: Jefe de producción agrícola
Dirección de empresa: Carretera Panamericana Norte Km. 523 Lote 10.7 Int. II – Virú, La Libertad

A quién convenga:

Con relación a los bachilleres Sanchez Otiniano Luis Fernando y Mendoza Vásquez Jhordyn Daniel, a quiénes se les otorgó una base de datos en formato Excel para uso exclusivo del desarrollo de la tesis con registros de producción y transporte de arándanos del año 2020 de la empresa Agroberries S.A.C. dónde me desempeñaba como jefe de producción agrícola, por tanto, se les hace mención en la presente carta que dicha información proporcionada no tiene permitida su divulgación a medios externos y se reitera que la usabilidad queda limitada hasta la finalización del trabajo de investigación.

Sin nada más que añadir, me despido.



.....
Diego Jesús Rodríguez Navarrete

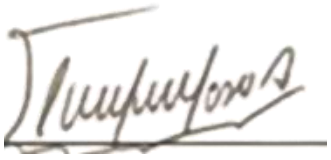
ANEXO 5. Constancia del asesor.

ACREDITACIÓN

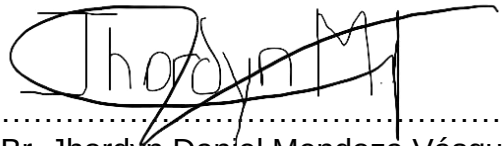
El **Dr. Walter Aurelio Lazo Aguirre**, que suscribe, asesor de la Tesis con Título **"MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO COMO SOPORTE EN LA TOMA DE DECISIONES EN EL ÁREA DE PRODUCCIÓN Y TRANSPORTE DE ARÁNDANOS EN LA EMPRESA AGROBERRIES S.A.C. – LA LIBERTAD 2022"**, desarrollado por los Br. En Ingeniería de Sistemas: Mendoza Vásquez y Sanchez Otiniano, acredita haber realizado las observaciones y recomendaciones pertinentes, encontrándose expedita para su revisión por parte de los señores miembros del Jurado Evaluador.

Trujillo, 2022

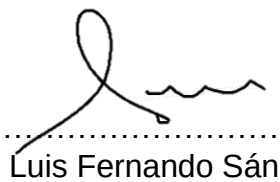
El Asesor:



……
Dr. Walter Aurelio Lazo Aguirre



……
Br. Jhordyn Daniel Mendoza Vásquez



……
Br. Luis Fernando Sánchez Otiniano