



**UNIVERSIDAD ESAN**

**FACULTAD DE INGENIERÍA**

INGENIERÍA INDUSTRIAL Y COMERCIAL  
INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

**Propuesta de un modelo de machine learning para el pronóstico de la demanda de prendas de vestir en la Corporación Brusko S.A.C.**

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial

Obtener el título profesional de Ingeniero de Tecnologías de Información y Sistemas

**AUTORES**

Ronald Alberto Ccoyccosi Choque

Luis Ernesto Huanay Palomino

Ethel Diana Huayllasco Chafloque

Vanessa Loayza Díaz

Katiuska Fiorela Mayorga Lopez

**ASESOR**

Junior Fabian Arteaga

ORCID N° 0000-0001-9804-7795

Diciembre, 2021

## ÍNDICE DE CONTENIDOS

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA .....	12
1.1. Descripción de la Realidad Problemática.....	12
1.2. Justificación de la Investigación .....	14
1.2.1. Teórica .....	15
1.2.2. Práctica.....	15
1.2.3. Metodológica .....	15
1.3. Delimitación de la Investigación.....	15
1.3.1. Espacial .....	15
1.3.2. Temporal .....	16
1.3.3. Conceptual .....	16
CAPÍTULO II: MARCO TEÓRICO .....	17
2.1. Antecedentes de la Investigación .....	17
2.2. Bases Teóricas.....	23
2.2.1. Inteligencia artificial .....	23
2.2.2. Machine Learning .....	26
2.2.3. Aprendizaje Supervisado .....	30
2.2.4. Demanda .....	37
2.2.5. Pronóstico .....	37
CAPÍTULO III: ENTORNO EMPRESARIAL.....	38
3.1. Descripción de la empresa.....	38
3.1.1. Reseña histórica y actividad económica .....	38
3.1.2. Descripción de la organización.....	39
3.1.3. Datos generales estratégicos de la empresa .....	41
3.2. Modelo de negocio actual (CANVAS) .....	45
3.3. Mapa de procesos actual .....	48
CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN .....	49
4.1. Diseño de la Investigación .....	49
4.1.1. Enfoque de la investigación .....	49
4.1.2. Alcance de la investigación .....	49
4.1.3. Diseño o tipo de la investigación.....	49
4.1.4. Población y muestra.....	49

4.2.	Metodología de implementación de la solución.....	50
4.3.	Metodología para la medición de resultados de la implementación .....	52
4.4.	Cronograma de actividades y presupuesto .....	53
CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN .....		54
5.1.	Propuesta solución.....	54
5.1.1.	Planteamiento y descripción de Actividades .....	54
5.1.2.	Desarrollo de actividades. Aplicación de herramientas de solución.....	55
5.2.	Medición de la solución. ....	59
5.2.1.	Análisis de Indicadores cuantitativo y/o cualitativo. ....	60
5.2.2.	Simulación de solución. Aplicación de Software .....	61
CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES.....		77
6.1.	Conclusiones .....	77
6.2.	Recomendaciones.....	77
REFERENCIAS BIBLIOGRÁFICAS.....		79
ANEXOS .....		83

## ÍNDICE DE FIGURAS

Figura 1.	Flujo del modelo propuesto.....	21
Figura 2.	Proceso de construcción de un modelo de Machine Learning.....	28
Figura 3.	Descripción genérica del flujo de entrenamiento de un algoritmo supervisado ...	31
Figura 4.	Gráfica de predicción .....	32
Figura 5.	Esquema de una red perceptrón multicapa con una capa oculta.....	33
Figura 6.	Transformación del espacio de entradas en un espacio de características .....	33
Figura 7.	Estructura de un árbol de decisión .....	34
Figura 8.	Representación gráfica de aprendizaje supervisado de un algoritmo de clasificación .....	35
Figura 9.	Representación gráfica del aprendizaje supervisado de regresión lineal .....	36
Figura 10.	Curva de la oferta y demanda.....	37
Figura 11.	Organigrama.....	39
Figura 12.	Cadena de suministro .....	40
Figura 13.	Canvas.....	47
Figura 14.	Mapa de procesos de Corporación Brusko S.A.C.....	48
Figura 15.	Diseño de la investigación .....	51
Figura 16.	Estructura de la base de datos de la compañía .....	55
Figura 17.	Extracción de reportes de pantalones de caballero .....	56
Figura 18.	Gráfico de dispersión del año 2018-2021 .....	66
Figura 19.	Gráfico de dispersión del año 2018.....	67
Figura 20.	Gráfico de dispersión del año 2019.....	67
Figura 21.	Gráfico de dispersión del año 2020.....	68
Figura 22.	Gráfico de dispersión del año 2021 .....	68
Figura 23.	Gráfico de dispersión del año 2018-2021 sin valores outliers .....	69
Figura 24.	Gráfico de dispersión en 3D.....	70
Figura 25.	Gráfico tipo botplox con valores outliers.....	70

Figura 26. Gráfico tipo botplox sin valores outliers.....	71
---	----

## ÍNDICE DE TABLAS

Tabla 1.	Exportaciones textiles .....	13
Tabla 2.	Conceptos de inteligencia artificial con el enfoque de diferentes autores .....	24
Tabla 3.	Matriz FODA .....	43
Tabla 4.	Matriz FODA cuantitativo .....	44
Tabla 5.	Población y muestra de la investigación .....	49
Tabla 6.	Comparación de metodologías.....	50
Tabla 7.	Metodología para medición de resultados.....	52
Tabla 8.	Cronograma de actividades.....	53
Tabla 9.	Presupuesto .....	53
Tabla 10.	Reporte de detalle de la demanda del producto .....	57
Tabla 11.	Reporte de ventas semanales pantalones caballero en Cusco .....	58
Tabla 12.	Cuadro comparativo de métricas de regresión .....	60
Tabla 13.	Dataframe de los años 2018-2021 .....	62
Tabla 14.	Dataframe del año 2019 .....	63
Tabla 15.	Tabla de los valores estadísticos 2018-2021 .....	63
Tabla 16.	Tabla de los valores estadísticos del año 2018.....	64
Tabla 17.	Tabla de los valores estadísticos del año 2019.....	64
Tabla 18.	Tabla de los valores estadísticos del año 2020.....	65
Tabla 19.	Tabla de los valores estadísticos del año 2021.....	65
Tabla 20.	Tabla de los valores estadísticos de los años 2018-2021 sin valores outliers .....	66
Tabla 21.	Cuadro comparativo de las ventas actuales vs ventas predichas 2018-2021 .....	73
Tabla 22.	Cuadro comparativo de las ventas actuales vs ventas predichas del año 2018.....	73
Tabla 23.	Cuadro comparativo de las ventas actuales vs ventas predichas del año 2019.....	74
Tabla 24.	Cuadro comparativo de las ventas actuales vs ventas predichas del año 2020.....	74
Tabla 25.	Cuadro comparativo de las ventas actuales vs ventas predichas del año 2021 .....	75

Tabla 26. Cuadro comparativo de las ventas actuales vs ventas predichas 2018-2021 sin valores outliers .....	75
--	----

## RESUMEN

En el Perú existen muchas empresas a cargo de la compra y venta de bienes (retails), para ellas la determinación de la demanda es crucial ya que esta impacta significativamente en los costos e ingresos que pueden llegar a tener. La presente investigación plantea pronosticar la demanda de los pantalones de caballero de la Corporación Brusko bajo la metodología CRISP-DM y empleando la técnica de regresión lineal. Se trabajó con datos brindados por la empresa (2018 – 2021), con ello se realizó el modelo de pronóstico y se obtuvieron los siguientes RMSE: 23.78 para el año 2018, 13.22 para el año 2019, 47.12 para el año 2020 y 17.87 para el año 2021. Además, se realizó el mismo análisis para la totalidad de años teniendo como resultado un RMSE de 59.07. Los datos presentaron outliers debido a la pandemia; eliminando estos datos atípicos se volvió a correr el modelo para la totalidad de los años obteniendo un RMSE de 29.98, el cual fue mejor comparado al modelo con la totalidad de datos sin outliers. Realizando el análisis de los resultados nos quedamos con el modelo para el año 2019, un año que no presenta outliers y con un adecuado RMSE.

Palabras claves: Demanda, pronóstico, crisp-dm, python, regresión lineal.

**ABSTRACT**

In Peru there are many companies in charge of the purchase and sale of goods (retails), for them the determination of demand is crucial since it significantly impacts the costs and income they may have. This research proposes forecasting the demand for men's pants from the Brusko Corporation under the CRISP-DM's methodology using the linear regression technique. We worked with data provided by the company (2018 - 2021) to forecast a model that better suits. The the following RMSE were obtained: 23.78 for 2018, 13.22 for 2019, 47.12 for 2020 and 17.87 for 2021. In addition, the same analysis was carried out for all years, resulting in a RMSE of 59.07. The data presented outliers due to the pandemic; removing atypical data, the model was run again for all the years, obtaining a RMSE of 29.98, which was better compared to the model with all data without outliers. After analyzing the results, the best model is the forecast from 2019, a year that does not present outliers and has an adequate RMSE.

**Keywords:** Demand, forecast, crisp-dm, python, linear regression.

## INTRODUCCIÓN

La industria textil es un mercado que está en constante crecimiento y en donde permanecer competitivo es una lucha constante. La competencia desleal con los productos provenientes de Asia que introducen sus productos al mercado peruano con precios muy por debajo de los que manejan las empresas locales y los factores externos, como políticos, económicos, sociales, naturales, etc. hacen que las empresas incorporen estrategias que les garantice ser siempre una buena opción para sus clientes.

Por lo cual se hace importante que las empresas apliquen herramientas que le permitan mejorar sus procesos, siendo una de ellas el pronóstico de la demanda, la cual va facilitar la realización de la planeación operativa y el presupuesto; por tanto, todo ello repercutirá directamente en el servicio al cliente. Para poder obtener buenos resultados es importante que se realice el pronóstico con la mayor precisión posible.

Hoy en día, la cantidad de datos que una empresa tiene a su disposición va a superar la capacidad que tienen las personas de poder asimilar, interpretar y tomar decisiones basados en esos datos, por lo cual la inteligencia artificial va a permitir que se puedan realizar esas tareas complejas a través de sus algoritmos.

Con la tecnología aprendida y accesible, y con información de la empresa Brusko S.A.C, los cuales realizaban el pronóstico de su demanda tomando de referencia su data histórica y la experiencia que han tenido por los años trabajados en el rubro, nos motivó a realizar el trabajo de investigación con el objetivo de contribuir con las MYPES para que puedan desarrollar este tipo de tecnología con la finalidad de mejorar sus procesos.

Esta investigación va a plantear el uso de algoritmo de aprendizaje supervisado de regresión lineal para pronosticar la demanda de la empresa Brusko S.A.C.

Por lo que la investigación fue dividida en seis capítulos, y cada capítulo va a ser enfocado con un lineamiento específico.

En el capítulo I se detalla el planteamiento del problema, la cual contiene una breve descripción de la realidad problemática, la justificación teórica, práctica y metodológica, finalizando con la delimitación del estudio de manera espacial, temporal y conceptual.

En el capítulo II se detalla el marco teórico, la cual contiene los antecedentes del caso y las bases teóricas para la estructura de la investigación.

En el capítulo III se detalla el entorno empresarial, el cual contiene la descripción de la empresa, el modelo de negocio de la empresa (canvas) y su actual mapa de procesos.

En el capítulo IV se detalla la metodología de la investigación, la cual contiene el diseño, tipo, enfoque, población y muestra del estudio, operacionalización de las variables y las técnicas para la recolección y procesamiento de datos. Además, se hará uso de la metodología CRISP-DM y se detalla el algoritmo y métricas usadas para la solución de la investigación.

En el capítulo V se detalla el desarrollo de la solución, en la cual vamos a presentar la descripción y desarrollo de las actividades, así como el análisis de los indicadores y la simulación del modelo usado.

Finalmente, en el capítulo VI se presentan las conclusiones y recomendaciones de la investigación, las cuales pueden ser tomadas en cuenta para futuras investigaciones o trabajos a realizar. Además, se incluye como anexo la entrevista con el especialista de la empresa Brusko.

## **CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA**

En el presente capítulo detallaremos la realidad problemática, la justificación y la delimitación de la investigación, los cuales indicarán de qué se trata la investigación, los problemas que presentan diferentes empresas tanto internacionales como nacionales, los beneficios que traerá la presente investigación y el campo en que se va a trabajar.

### **1.1. Descripción de la Realidad Problemática**

Según Garcete, Benítez, Pinto-Roa y Vazquez (2017) determinar la demanda de las ventas es una de las tareas más complicadas de todos los negocios de retail. La naturaleza cambiante del mercado en sí hace que esta varíe constantemente y las empresas busquen formas de acercarse a su valor real mediante diferentes técnicas.

El sector Retail, que se encarga de la venta de sus productos al cliente final, busca mantener constantemente su nivel de inventarios lo más cercano posible al nivel de la demanda para poder optimizar sus ganancias.

En el Perú, este sector ha estado en constante crecimiento, excluyendo el periodo de confinamiento por el que pasó el país. Según GFK Perú, el crecimiento ha sido de 88% en este primer trimestre con respecto al 2020 y en un 50% comparándola con años pasados.

Según El Comercio (2021), la Cámara de Comercio de Lima (CCL) hace referencia que *las exportaciones textiles peruanas registraron un crecimiento entre 10% a 15% para el cierre del año con respecto a lo reportado en el 2019*, tal como se puede visualizar en la tabla 1.

Tabla 1. Exportaciones textiles

PAÍS	ENE-JUN 2020	ENE-JUN 2021	VARIACIÓN %	PARTICIPACIÓN %
Total exportado	388.778.579	725.835.824	86,70	100,00
EE.UU.	207.415.603	403.933.046	94,75	55,65
Unión Europea	34.680.518	55.838.287	61,01	7,69
Chile	17.016.173	40.542.564	138,26	5,59
Colombia	14.824.648	30.704.526	107,12	4,23
Canadá	9.956.621	29.668.337	197,98	4,09
Brasil	16.869.777	24.438.414	44,87	3,37
Ecuador	13.957.194	22.234.227	59,30	3,06
China	14.041.887	19.741.115	40,59	2,72
Bolivia	9.836.899	16.675.559	69,52	2,30
Reino Unido	6.596.764	13.064.386	98,04	1,80
Demás países	43.582.494	68.995.363	58,31	9,51

Fuente: El Comercio (2021)

Según los reportes de las encuestas de Ernst & Young (EY) Perú en el 2018, el 75% de los negocios en el país han tenido como principal tema de conversación la automatización de procesos e inteligencia artificial en sus directorios.

La empresa peruana Brusko S.A.C, retail de prendas de vestir que opera a nivel nacional tiene como uno de los principales problemas la determinación de la demanda. El área comercial

tiene el gran reto de satisfacer la demanda de los clientes; actualmente la empresa hace una estimación empírica de la demanda en base a experiencia y data histórica. Sin embargo, esto no siempre es suficiente debido a que se tiene presencia en distintas regiones del país, las cuales tienen diferencia en gustos, tiempos de compra, estacionalidad, y otros.

Hay temporadas en las que el área comercial reporta pérdidas de venta por falta de abastecimiento en ciertos productos, los cuales son enviados desde el almacén siempre y cuando se tenga en stock, sino el producto requerido tiene que comprarse o mandarlo hacer (esta etapa es tercerizada) para después poder ser enviado a las tiendas que lo requieran, lo cual toma entre 2 a 3 semanas. Estas pérdidas se dan ya que la empresa no cuenta con una metodología o modelo que les permita anticipar la demanda estimada.

El área de compras hace lo posible por satisfacer las solicitudes de los puntos de venta; sin embargo, no cuenta con un plan de adquisición establecido para cubrir la cantidad demanda por punto de venta. Su enfoque actual se basa en compras de gran volumen para obtener mejores precios, lo cual genera volumen de inventario de los productos con baja rotación y esto generaría costos adicionales para la empresa.

Los factores externos, la pandemia y el sector político, que han afectado al mercado peruano en los últimos años han provocado un alza en los precios de productos, disminución en las ventas, incremento del costo de operaciones y otros que hacen necesario un cambio en el modelo empírico actual para la estimación de la demanda hacia una automatización del proceso.

De seguir sin ninguna actualización en los procesos actuales, Brusko seguirá perdiendo oportunidades de venta y con el tiempo dejará de ser tomado como una opción de compra para sus clientes. En esta investigación proponemos el uso de la técnica de machine learning para predecir la demanda en prendas de vestir de pantalones de caballero para la empresa Brusco S.A.C.

## **1.2. Justificación de la Investigación**

De acuerdo a Mendez (1995), la justificación de la investigación presenta tres criterios: Teórica, Práctica y Metodológica; criterios que se justificarán en los siguientes puntos.

### **1.2.1. Teórica**

La presente investigación servirá para ayudar a que la Corporación Brusko pueda pronosticar la demanda de los pantalones con el fin de aprovechar las oportunidades de venta en todo momento.

Con este trabajo se va proponer la implementación de pronóstico de la demanda mediante el uso del modelo de regresión lineal de Machine Learning con el fin de reducir las pérdidas en ventas por falta de stock.

Además, este trabajo pretende dejar una base para que sea evaluado y discutido por la empresa para su implementación ya que actualmente no cuentan con algo parecido.

### **1.2.2. Práctica**

La presente investigación busca a través del modelo regresión lineal de machine learning pronosticar la demanda de la categoría de productos de pantalones, usando la herramienta Python y con ello alertar al área comercial y al área de abastecimiento la correcta planificación en la reposición del stock en cada local, disminuyendo el trabajo manual, trabajo que realiza el encargado de tienda 1 vez por semana, enviando la solicitud de reposición.

De igual forma, disminuir los tiempos largos de reposición, tiempos que pueden ser hasta 4 semanas desde el proceso de compra hasta el despacho a las tiendas. Con ello esperamos eliminar los trabajos manuales y disminuir los tiempos de reposición hasta en un 30%.

### **1.2.3. Metodológica**

En este punto se busca que el modelo propuesto, sea implementado en tres etapas donde se abarque el tratamiento de la data, la modelación para el pronóstico deseado y la validación respectiva del modelo. Estas etapas serán explicadas más adelante en la presente investigación.

## **1.3. Delimitación de la Investigación**

### **1.3.1. Espacial**

El presente estudio se realiza para la Corporación Brusko, empresa retail de ventas de prendas de vestir ubicada en Perú y con tiendas distribuidas en varios departamentos del país. Como parte del estudio se tomará los datos de venta de pantalones de caballero en las sucursales

de Cusco que sean proporcionadas por la empresa. No se tomará en cuenta las variables externas que afectan a la demanda que no sean proporcionadas por la empresa.

### **1.3.2. Temporal**

Cabe indicar que se trabajará con la información proporcionados por la empresa; entre ellos se recolectarán datos reales desde el 2018 para poder calcular la demanda, estos datos nos darán información sobre las ventas por establecimiento que permita pronosticar la demanda y así poder reponer a tiempo la mercadería en las tiendas lo cual permitirá que no se pierdan ventas.

### **1.3.3. Conceptual**

El presente estudio se centrará en el pronóstico de la demanda de la empresa Brusko basado en el modelo de regresión lineal que se ha visto en el curso de actualización de machine learning, el cual tiene como objetivo determinar las ventas a futuro para poder conocer la cantidad de pantalones que se requieren para reponer en las distintas tiendas que tiene la compañía.

## CAPÍTULO II: MARCO TEÓRICO

En el presente capítulo se explicarán los antecedentes encontrados y las bases teóricas usadas, las cuales ayudarán a entender de una mejor manera el problema a tratar. Además, se evaluarán los resultados previos de otras investigaciones relacionadas al presente trabajo.

### 2.1. Antecedentes de la Investigación

**Céspedes, A. (2017), en Santiago, Chile, realizó la investigación titulada: *Construcción De Modelo De Forecast Para Estimación De Demanda En Una Empresa Multinacional De Retail*, Memoria De Titulación Para Optar Al Título De Ingeniero Civil Informático.**

El autor realiza este trabajo sobre la empresa Mars sede Chile, empresa de dulces multinacional de origen estadounidense, que fue creada en el año 1911 y que cuenta con más de 72000 empleados en todo el mundo, más de 20 plantas, presencia en más de 78 países y con ventas netas de más de 33 mil millones de dólares.

La empresa cuenta con 5 segmentos comerciales y en Chile solo comercializa 2, Mars Chocolate que tiene como marcas a Snickers, M&Ms, Milky Way y Twix, Mars Petcare que tiene como marcas a Pedrigee y Whiskas. Estos productos se importan desde Estados Unidos, México y Argentina.

Dentro de los problemas detectados por el autor en la empresa, indica que los tiempos de importación son muy extensos y que pueden demorarse hasta 3 meses el proceso de importación desde Estados Unidos. Esto les genera un problema ya que los chocolates tienen una vida útil de 12 meses y cuando llegan a Mars Chile tienen una vida útil de 9 meses, y los supermercados y mayoristas aceptan productos con una vida útil de 6 meses, dejando entre 2 a 3 meses a la empresa para que los pueda comercializar. Como segundo problema, indica que la demanda no es regular y hay periodos de sobre stock y quiebres de stock por cambios en el mercado, y que es necesario contar con un inventario de seguridad.

La empresa tiene como objetivo mejorar la predicción de la demanda para manejar un forecast accuracy del 70% como mínimo, ya que actualmente llegan entre 55% a 60%, lo que equivale a 15 puntos porcentuales por debajo del objetivo. La predicción se realizará teniendo en cuenta la historia de venta de cada producto.

Para poder predecir la demanda utiliza archivos en texto plano descargados desde SAP y archivos maestros manejados por el área comercial y marketing. Entre estos archivos se tienen el histórico de la demanda de los años 2018 al 2020, histórico de productos actuales y discontinuados, e histórico de clientes de los años 2018 al 2020.

Las variables que tiene el Histórico de la demanda son:

- Fecha de compra, canal de compra, código del cliente, razón social del cliente, ruc del cliente, tipo de producto, SKU del producto, nombre del producto y demanda del producto en unidades.

Las variables que tiene el histórico de clientes son:

- Ruc del cliente, código del cliente, razón social del cliente, región, área, tipo de cliente, código de vendedor y nombre del vendedor.

Las variables que tiene el histórico de productos son:

- SKU del producto, nombre del producto, tipo de producto, segmento del producto, unidad de medida, identificador único del producto, identificador único del envase, estado del producto.

De las variables mencionadas, se trabajarán con 4 de ellas para el pronóstico de la demanda. Estas variables son: fecha de compra, demanda del producto, cliente y producto.

Para el desarrollo de la predicción, el autor usa como técnica las Redes Neuronales Artificiales (RNA), ya que según su investigación es una técnica muy usada, que cuenta con una importante habilidad de aprendizaje y brinda respuestas en tiempo real. Para trabajar con RNA usa el lenguaje R, ya que es una herramienta enfocada al análisis estadístico y cuenta con varias bibliotecas con modelos predictivos. En el presente trabajo el autor usa la biblioteca forecast.

Como resultado, el autor llega a las siguientes conclusiones:

- Anteriormente Mars realizaba predicciones sin evaluar la calidad y veracidad de los datos y con el nuevo modelo se tienen una predicción más real y cercana a la compañía y permite tener una mejor calidad de los datos, así como automatizar de manera correcta los procesos de predicción.

- La etapa de compresión de datos es muy importante y fue difícil de desarrollar, ya que se obtuvieron datos de distintas áreas y se tuvo que trabajar en el relacionamiento de estas.
- Se hicieron varios análisis y pruebas usando el modelo de RNA, esto para poder tener énfasis en el objetivo principal del proyecto y se consiguió tener un resultado satisfactorio, llegando a tener un 80% de forecast accuracy.

**Garcete, A., Benítez, R., Pinto-Roa, D. y Vázquez, A. (2017), Argentina, realizó la investigación titulada: *Técnica de pronóstico de la demanda basada en Business Intelligence y Machine Learning*. Simposio Argentino sobre Tecnología y Sociedad (STS)**

Los autores en su artículo hablan sobre la gestión de la demanda en el sector retail, y que uno de los principales problemas es enfrentar el manejo eficiente del stock, problema que se debe evitar, ya que, el exceso de stock incurre en sobrecostos de almacenaje o la falta de dichos productos conlleva a pérdidas de ventas y a la insatisfacción de cliente. Así mismo indican, que las técnicas de pronóstico pueden ser cuantitativas o cualitativas y que de igual forma el problema sigue siendo la falta de confiabilidad del modelo de predicción. Por lo que se tiene una interrogante, la cual es si la demanda predecida será superior o inferior a la demanda real.

Para poder llevar a cabo el ejercicio cuentan con una base de datos relacional, con data histórica del 2018 al 2020, con las operaciones de una empresa retail que comercializa artículos alimenticios y artículos de limpieza. La base de datos contiene las siguientes tablas: productos, proveedores, ventas cabeceras, ventas detalle y movimiento de stock. Con dicha data aplican un proceso de ETL para luego generar un Datawarehouse con las tablas de hechos y dimensiones. Luego de ello generan una nueva salida, en texto plano, de datos solo con las variables para ser usado en el modelo de pronóstico y que muestran la información de ventas de forma semanal, quincenal y mensual de manera clasificada.

Los autores también brindan una definición de machine learning, categorías de algoritmos y tipo de problemas, tales como: regresión, clasificación, segmentación y análisis de red. Así mismo, mencionan técnicas de evaluación de resultados, tales como: cross-validation y Stratified k-fold Cross-Validation.

Para la fase de procesamiento del dataset, usan la herramienta WEKA, software de machine learning basado en java, para clasificar usan ZeroR, ya que es y para la evaluación de resultados usan Stratified k-fold Cross Validation con un k igual a 10.

Los autores concluyeron que sería conveniente que los expertos en compras establezcan sus propios criterios para el modelo de predicción. Los periodos evaluados son los comunes para el área comercial y en la práctica los periodos se toman de manera estratégica.

Así mismo indican, que en relación a los resultados experimentales se obtuvieron altas tasas de acierto, las cuales superan el 98%, y que la técnica propuesta pretende que el nuevo modelo sea una herramienta de apoyo en la toma de decisiones de compras para la reposición de stock.

**Pérez, R. (2019), Lima, Perú. *Generación de reglas de asociación para productos de retail utilizando el algoritmo FP-Growth paralelo*. Universidad de Lima.**

Existe una tendencia de registrar y almacenar toda data posible relacionada a procesos operativos de una organización. En el sector retail gran proporción de esta data proviene de las infinitas transacciones efectuadas en cada sucursal

Empresas representativas a nivel mundial del sector han identificado el potencial en el hecho de minar la información oculta dentro de estos conjuntos de datos. No obstante, no todas las organizaciones tienen como estándar la aplicación de las técnicas de minería de datos. En tal sentido, la mayor aplicación de la minería de datos en el sector retail es la minería de reglas de asociación. Esta práctica supone un atractivo para las organizaciones debido a que tiene la intención de convertir la gran cantidad de datos que poseen en información útil y conocimiento valioso para la toma de decisiones como, por ejemplo: el diseño de la cesta de productos, la elección de la estrategia de promoción.

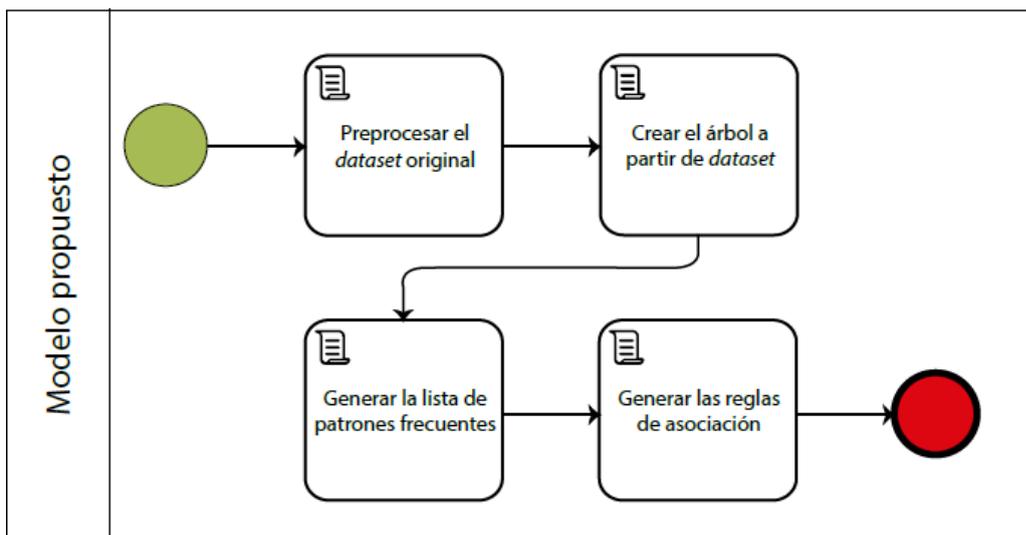
El autor realiza la investigación en un comercio minorista con sucursales ubicadas en Lima, Perú. El conjunto de datos utilizado en esta investigación corresponderá a las transacciones realizadas con el tiempo para productos de consumo masivo y se emplea el algoritmo FP-Growth bajo el marco teórico de Apache Spark. El algoritmo de FP-Growth busca resolver las desventajas que posee el algoritmo clásico, ya que este consiste en dos pasos para la generación de las reglas de asociación, además de ser menos costoso en tiempo y en recursos.

El autor realizó inicialmente una limpieza al dataset para luego realizar un análisis exploratorio de la data y con ello tener conocimiento de cuál es el comportamiento de las variables. Para efectos de la investigación, sería la relación entre productos y departamentos o categorías, los diez productos con más órdenes, las horas o días de la semana con más órdenes, etcétera.

Se empleó Python ya que este es el lenguaje más usado en Data Science. Lo primero que se realizó fue identificar los módulos necesarios para la solución. Estos son, un módulo para el preprocesamiento y exploración de la data y otro módulo para el algoritmo de generación de ítems frecuentes y generación de reglas de asociación. Este flujo se puede observar en la figura 1.

El modelo se inicializó con los valores 0,01 para soporte mínimo y 0,1 para confianza. Los ítems frecuentes encontrados que sobrepasen el umbral de soporte están conformados por dos ítems. Referente a las reglas de asociación, la relación más fuerte es entre palta Hass orgánica (antecedente) y bolsa de plátanos orgánicos (consecuente). Esta relación se refiere a la probabilidad de que ocurra el consecuente a partir del antecedente. El autor recomienda que hay que considerar que las necesidades del negocio y sus objetivos estratégicos juegan un papel dominante y crítico en la implementación de un modelo como el propuesto en este trabajo (figura 1). Este análisis define los parámetros de funcionamiento del modelo y los resultados obtenidos tienen que ser relevantes y consistentes con la información del negocio.

Figura 1. Flujo del modelo propuesto



Fuente: Pérez, R. (2019).

**Cano, J. (2021), Colombia. *Desarrollo de Solución Analítica para la predicción de la demanda de línea*. Universidad de Antioquia.**

Una de las situaciones problemáticas que tiene que afrontar y aprender una organización que comercializa bienes como lo son los alimentos, es el fenómeno conocido en mercadeo como “canibalización”. Este fenómeno como se profundizará posteriormente afecta a las organizaciones en la medida de que la oferta o promoción de uno de sus productos no solo afecta a su competencia sino también a ella misma, ya que, a pesar del aumento de la demanda del producto de oferta, consecuentemente se genera la disminución de uno o más productos de línea de su portafolio. Debido a esto, las empresas tienen la necesidad de desarrollar nuevas herramientas que les permita conocer el comportamiento de la nueva demanda y cuál es la relación entre los productos de oferta y los de línea.

El autor desarrolló el presente trabajo con el objetivo de proponer un modelo de predicción de ventas, que permite estimar de la manera más asertiva el comportamiento de las ventas del producto de línea cuando se tiene la presencia de una oferta, tomando como punto de partida, la realidad y contexto de una empresa manufacturera de alimentos cárnicos de Medellín. Este modelo está basado en la predicción utilizando Machine Learning para un solo producto de oferta que afecta a un solo producto de línea en un perfil (cliente) específico. La metodología implementada fue el framework CRISP-DM, específico para la ciencia y minería de datos.

El modelo CRISP-DM se sigue utilizando y está basada en un modelo de proceso jerárquico que consta de seis fases, algunas de las fases son bidireccionales, lo que significa que es posible devolverse en el proceso para mejorar o resolver inconvenientes, por lo cual la sucesión de fases no tiene que ser ordenada. Las fases del modelo son:

01. Entendimiento del negocio: En esta fase se identificaron y definieron las variables, las cuales fueron evaluadas y de acuerdo con su disponibilidad, entendimiento y requerimientos serán seleccionadas o no como insumo para el ejercicio.
02. Entendimiento de los datos: Para esta fase se usó SQL Server para procesar la data de historial de ventas de la referencia de oferta y las ventas de la referencia de línea correspondiente por semanas por un periodo de dos años (2018 y 2019). En esta etapa se pudo identificar también que los perfiles con mayor cantidad de demanda de estos productos, se destacan el canal “AUT INDEPENDIENTE”, “EXITO” y “OLIMPICA”.

03. Preparación de los datos: El perfil seleccionado para realizar un primer modelo exploratorio fue el perfil “ÉXITO”, el cual fue seleccionado por que tiene una gran cantidad de eventos con respecto a otros perfiles.
04. Modelamiento: Una vez obtenido el DataSet final preparado, se procede a realizar la técnica para modelación “train-test split” o “separación de entrenamiento y prueba”. Para ello, se obtuvo un conjunto de entrenamiento de 72 datos (80%) y 18 de prueba (20%). El tipo de modelación que se le realizó a los datos fue Deep learning con redes neuronales artificiales.
05. Evaluación: En esta fase, se evaluó la métrica de rendimiento R2 o coeficiente de determinación para el conjunto de prueba, el cual fue de 0.7856, el cual es un resultado sobresaliente con respecto a otras modelaciones.
06. Implementación: Finalmente, se establecieron relaciones en el transcurso de un periodo de 6 meses con la unidad de analítica avanzada de una sede del grupo empresarial que presta servicios de tecnología y con ella se crearon acuerdos para continuar la industrialización e implementación de esta solución a escala para su posterior uso real. Esto quiere decir que, esta fase no se llevó a cabo totalmente.

El autor afirma que se logró aplicar técnicas de exploración de datos para analizar y clasificar las variables más significativas para el entendimiento del problema, obteniendo ideas y conclusiones importantes a tener en cuenta en el desarrollo de la transformación y modelado de datos. Asimismo, se evaluaron diferentes técnicas de Machine Learning, con el fin de obtener ideas y resultados de valor acerca del fenómeno de canibalizaciones, cómo se evidencio en el presente informe, por simplicidad se observa el desarrollo de la modelación con mejor resultado general y global.

## **2.2. Bases Teóricas**

### **2.2.1. Inteligencia artificial**

Antes de definir a profundidad Inteligencia Artificial, primero se detallará el significado de cada una por separado. Según La Real Academia de la Lengua Española, la inteligencia es la capacidad de poder conocer, comprender o entender y artificial es algo que fue producido por el ingenio del ser humano. Por lo cual se detalla a continuación las definiciones más relevantes, las cuales se pueden visualizar en la tabla 2 (Russel&Norvig, 2008).

Las definiciones que se encuentran en el lado izquierdo de la tabla miden el éxito de la forma que actúan los seres humanos, el cual debe incluir hipótesis y una constatación mediante experimentos, por el contrario, en el lado derecho se toma como referencia un concepto de racionalidad, que consiste en que “si hace lo correcto, va a depender de su conocimiento”, así como también consiste en una combinación de ingeniería y matemática.

Tabla 2. Conceptos de inteligencia artificial con el enfoque de diferentes autores

<b>Sistemas</b>	<b>Como humanos</b>	<b>Racionalmente</b>
<b>Que piensan</b>	<p>“El nuevo y excitante esfuerzo de hacer que los computadores piensen ... máquinas con mentes, en el más amplio sentido literal”. (Haugeland, 1985)</p> <p>“[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje...” (Bellman, 1978)</p>	<p>“El estudio de las facultades mentales mediante el uso de modelos computacionales”. (Charniak y McDermott, 1985)</p> <p>“El estudio de los cálculos que hacen posible percibir, razonar y actuar”. (Winston, 1992)</p>
<b>Que actúan</b>	<p>“El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia” (Kurzweil, 1990)</p> <p>“El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor”. (Rich y Knight, 1991)</p>	<p>“La Inteligencia Computacional es el estudio del diseño de agentes inteligentes”. (Poole et al., 1998)</p> <p>“IA está relacionada con conductas inteligentes en artefactos” (Nilsson, 1998)</p>

Fuente: Russel, S. y Norvig, P. (2008).

A continuación, se va a describir los 4 enfoques que se han mencionado en la tabla anterior:

- **Pensar como ser humano: enfoque del modelo cognitivo**

Según Russel & Norvig (2008), para decir que un programa piensa como un ser humano, es indispensable tener un mecanismo que permita determinar la forma de pensar de los humanos, teniendo dos maneras de realizarlo: a través de introspección, que se trata de intentar atrapar nuestros pensamientos conforme van apareciendo y por medio de experimentos psicológicos. Si se tiene una teoría debidamente precisa en relación a cómo es que trabaja la mente, dicha teoría se puede expresar en un programa de computador. Si se tiene que los datos, tanto de entrada como de salida, del programa y los tiempos de respuesta son parecidos a los de una persona, se tiene la evidencia de que una proporción de los mecanismos de dicho programa se puede contrastar con los que desarrollan los humanos. En lo que respecta a la ciencia cognitiva, convergen modelos computacionales de inteligencia artificial y las técnicas experimentales de psicología procurando realizar teorías precisas y comprobables con respecto al funcionamiento de la mente del ser humano.

- **Comportamiento del ser humano: Prueba de turing**

Alan Turing fue el que propuso la prueba de Turing, la cual se diseñó para ofrecer un concepto operacional y satisfactorio de inteligencia. Para poder disponer de inteligencia artificial, Turing propuso una prueba que se basa en la “incapacidad de diferenciar entre entidades inteligentes indiscutibles y seres humanos” (Russel & Norvig, 2008, p. 3). El computador va a superar la prueba siempre y cuando un evaluador (humano) no sea capaz de poder distinguir si las respuestas proporcionadas de preguntas planteadas pertenecen a una persona o no. Según Russel y Norvig (2008) El computador tiene que poseer capacidades, los cuales se mencionan a continuación:

- Procesamiento de lenguaje natural
- Representación del conocimiento para poder almacenar todo lo que se conoce
- Razonamiento automático, se recurre de lo almacenado para poder hacer frente a las preguntas y poder sacar nuevas conclusiones
- Aprendizaje automático que le permita adaptarse a nuevas situaciones y para identificar y extrapolar patrones.

- **Pensar racionalmente: un enfoque de las leyes del pensamiento**

Aristóteles fue uno de los primeros filósofos en codificar “la manera correcta de pensar”, esto quiere decir un razonamiento que no se puede refutar. Sus silogismos son argumentaciones que mediante premisas adecuadas llevan a una conclusión idónea, por ejemplo “Si todos los hombres son mortales y Sócrates es un hombre entonces Sócrates es mortal” (Russel & Norvig, 2008).

- **Actuar racionalmente: un enfoque del agente racional**

Agente (del latín agere, que es hacer), viene a ser algo que razona, pero los agentes informáticos deben tener otros atributos que los puedan diferenciar de los programas convencionales, tales como que sean dotados de controles que sean autónomos, puedan percibir su entorno, que sigan en un lapso de tiempo prolongado, adaptación al cambio y que puedan lograr objetivos diferentes. Por lo cual un agente racional es quien actúa con el fin de lograr el mejor resultado y cuando se tiene incertidumbre, el óptimo resultado esperado. (Russel & Norvig, 2008)

### 2.2.2. Machine Learning

Según Bobadilla (2020) Machine learning (llamado en adelante ML) es la ciencia que hace que a partir de datos 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. El proceso integral de ML no es fácilmente automatizable ya que un ingeniero de datos tiene a su cargo varias tareas que son relevantes para el proceso, los cuales son:

- Identificar la fuente de datos y realizar su limpieza
- Eliminar información que es muy correlacionada
- Buscar información sesgada
- Realizar las normalizaciones que se necesitan
- Identificar los tipos de soluciones de ML cuya aplicación sea la adecuada
- Elegir el algoritmo correcto
- Ajustar los hiper parámetros del método que ha sido elegido
- Analizar los resultados
- Identificar comportamientos que no han sido correctos
- Corregir algún proceso anterior para cambiar lo que sea necesario y así poder tener mejores resultados.

### 2.2.2.1. Tipos de aprendizaje de machine learning

Para poder tratar cualquier tarea en concreto, el ingeniero de datos tiene que conocer conceptos relevantes de ML y las distintas opciones que existen. Según Bobadilla (2020) en los conocimientos básicos tenemos la identificación de tareas, que se empieza con clasificar los problemas de ML en uno de los siguientes tipos:

- Aprendizaje supervisado: cuando cada dato o un conjunto de datos de entrada tiene asociada una etiqueta.
  - Regresión: Por ejemplo, se puede predecir el precio de una casa teniendo en cuenta información como la cantidad de habitaciones y la distancia que tiene a la avenida principal.
  - Clasificación: Partiendo del ejemplo que se tiene un grupo de imágenes, en donde cada una contiene un tipo de metadato, una etiqueta o un conjunto de estas, la clasificación predice la etiqueta que le corresponde a una nueva imagen, que no está incluida en el grupo de datos
- Aprendizaje no supervisado: usa información no etiquetada.
  - Clustering (agrupamiento): su objetivo es agrupar muestras.
  - Reducción de dimensiones: se usa generalmente como una fase de pre-procesamiento de datos básicamente en regresión o clasificación. Por ejemplo, se trabaja para un sistema de recomendación.
- Aprendizaje semi-supervisado: trabaja con un grupo de datos, los cuales tiene un porcentaje de datos que están etiquetados y otros que no lo están.
- Aprendizaje por refuerzo: el algoritmo de aprendizaje va a recibir información de un escenario real o simulado, cuando el sistema efectúa una acción, esta es penalizada o recompensada.

### 2.2.2.2. Pasos para elaborar un modelo de ML

Crear un modelo de ML no solo es usar el algoritmo de aprendizaje también implica un proceso que detallaremos a continuación en la figura 2:

Figura 2. Proceso de construcción de un modelo de Machine Learning



Fuente: Manrique, E. (2020).

Se hace una descripción de cada uno de los pasos mencionados en la imagen anterior, los cuales se van a describir a continuación (Lopez, 2015, citado por Manrique, 2020):

- 1.- Recolección de datos: Los datos se pueden recopilar de una empresa o de un sitio web. Este paso es muy importante porque es el principio del cual se parte para lograr un resultado exitoso.
- 2.- Preprocesamiento de los datos: Se debe garantizar que todos los datos tengan un formato apropiado para poder alinear el algoritmo.
- 3.- Exploración de datos: Se hace un análisis para poder corregir si hay valores que faltan o encontrar un patrón que facilite la creación del modelo (características con mayor influencia para poder hacer la predicción).
- 4.- Entrenar el algoritmo: Los algoritmos deben alinearse con los datos (los que se procesaron en los procesos anteriores) para que los algoritmos extraigan información relevante y se pueda realizar las predicciones.
- 5.- Evaluación de los algoritmos: Se va a realizar una evaluación con respecto a la precisión del algoritmo con respecto a las predicciones que hizo. Si no se está de

acuerdo con los resultados se debe regresar a la etapa anterior y cambiar los parámetros necesarios para lograr el rendimiento deseado.

### **2.2.2.3. Lenguajes de desarrollo de ML**

Se tiene múltiples lenguajes de programación que se usan para realizar aplicaciones de inteligencia artificial y machine learning. Para nuestra investigación vamos a usar el lenguaje de programación Python el cual se describe a continuación:

#### **2.2.2.3.1. Lenguaje Python**

Python es un lenguaje donde su código se ejecuta en el navegador al cargar la página, es independiente de la plataforma y orientado a objetos, está listo para realizar cualquier tipo de programa, desde aplicaciones de Windows hasta servidores de red o incluso páginas web. Es un lenguaje interpretado, lo que ofrece ventajas como la velocidad de desarrollo e inconvenientes como una velocidad más baja al ser ejecutado. (Álvarez, 2003, citado en Manrique, 2020, p. 590)

A continuación, se nombrará algunas características de Python para la realización de ML:

- Python tiene una extensa biblioteca y macros los cuales ayudan en la codificación y se ahorra tiempo en su desarrollo.
- Tienen códigos concisos y legibles, los cuales facilitan su uso
- Es rápido en su desarrollo y no se necesita implementar el algoritmo para poder probarlo.
- Es fácil de interpretar

#### **2.2.2.3.2. Lenguaje R**

Es un lenguaje de programación creado para hacer análisis estadísticos y diseñar gráficos a partir de datos, tiene diferentes ventajas como que es un software libre, es multiplataforma.

Algunas características del lenguaje R son:

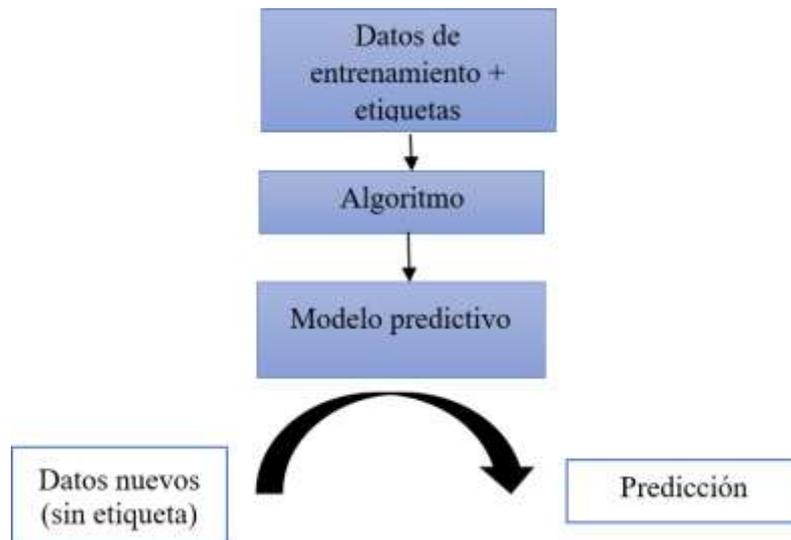
- Tiene bibliotecas preconstruidas a partir de las cuales los usuarios pueden definir sus propias funciones.

- Las funciones se pueden manejar como los vectores, se pueden guardar en listas y establecerlas a variables.
- Es integrable, puede modificar datos en diferentes tipos de bases de datos, además de poder interactuar con otros lenguajes como python o perl.
- Tiene librerías que permiten tener un gran número de gráficos desde los más simples a complejos.
- Está basado en la memoria, todos los objetos que se determinen en el programa quedan en la memoria de la máquina. (León, 2021)

### **2.2.3. Aprendizaje Supervisado**

El aprendizaje supervisado “se refiere a un tipo de modelo de machine learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida (outcome, eventos o labels) son conocidos” (Beunza, Puertas & Condés, 2019, p. 35). Estos modelos van aprendiendo de los resultados que se conocen y se realizan ajustes en los parámetros internos para poder ajustarse a los datos de entrada. Una vez teniendo el modelo debidamente entrenado y los parámetros interiores coherentes con los resultados que se tiene de los datos de entrenamiento y datos de entrada, el modelo va a poder realizar predicciones idóneas ante nuevos datos que no se han procesado anteriormente y de los que no se conoce el outcome, label o evento. En la figura 3 se podrá visualizar la descripción del flujo de entrenamiento del aprendizaje supervisado.

Figura 3. Descripción genérica del flujo de entrenamiento de un algoritmo supervisado



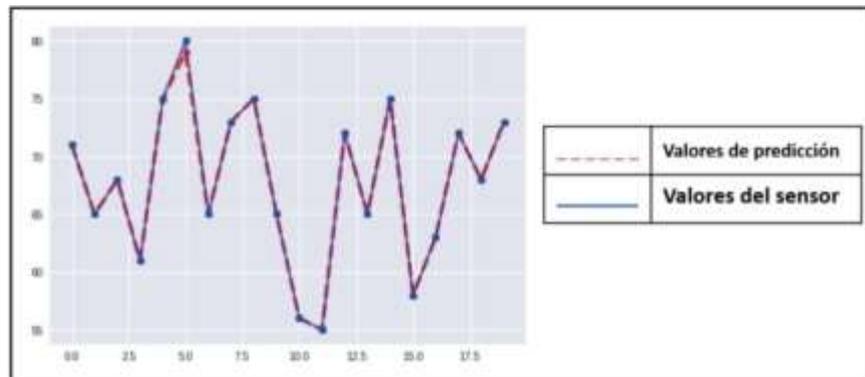
Fuente: Beunza, Puertas y Condés. (2019).

### 2.2.3.1. Modelos de aprendizaje supervisado

#### 2.2.3.1.1 Naive Bayes

Al tener que modelar sistemas inteligentes para su implementación en el mundo real, se va a tener que enfrentar con la incertidumbre, la cual se debe a la carencia de modelar todas las excepciones y condiciones que se tiene. Teoremas como el que conocemos como Naives Bayes va a permitir la construcción de algoritmos de predicción y de inferencia los cuales van a facilitar la identificación automática de las regularidades que se pueden encontrar en los datos, lo cual va a permitir aplicar acciones como la clasificación de datos en distintas categorías. Los métodos de agrupación de datos, los cuales son en su mayoría probabilísticos, se van a fundamentar en lo que llamamos la regla de Bayes, la cual va a permitir unir información de observación y previa. En la figura 4, con una efectividad de 95%, se va a mostrar los resultados de la predicción de los valores futuros de temperatura que se van a obtener de los datos que brindan los sensores (Sittón, Rodríguez y Muñoz, 2018).

Figura 4. Gráfica de predicción



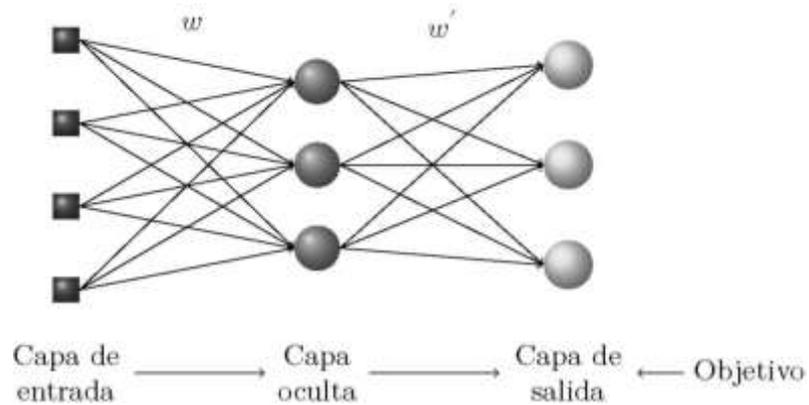
Fuente: Sittón, Rodríguez y Muñoz. (2018).

### 2.2.3.1.2 Neural Network

Imita las características de la mente para obtener la capacidad de procesamiento de la información. Es un sistema de neuronas que están interconectadas las cuales colaboran entre sí para generar un estímulo de salida. El proceso de información puede ser distribuida y paralela (Haykin, citado en González y Alba, s.f.).

Se puede aplicar redes neuronales tanto para un modelo de aprendizaje supervisado, en el cual se representa las relaciones complejas entre las variables de entrada y salida, como para un modelo no supervisado, donde se clasifica los datos de los cuales se desconoce a priori ninguna clase de organización. Una de las redes más idónea es la red perceptrón multicapa (figura 5) que contienen por lo menos una capa oculta con bastantes unidades no lineales, pudiendo aprender algún tipo de función o relación continua en medio de un grupo de variables tanto de entrada como de salida (González y Alba, s.f.).

Figura 5. Esquema de una red perceptrón multicapa con una capa oculta

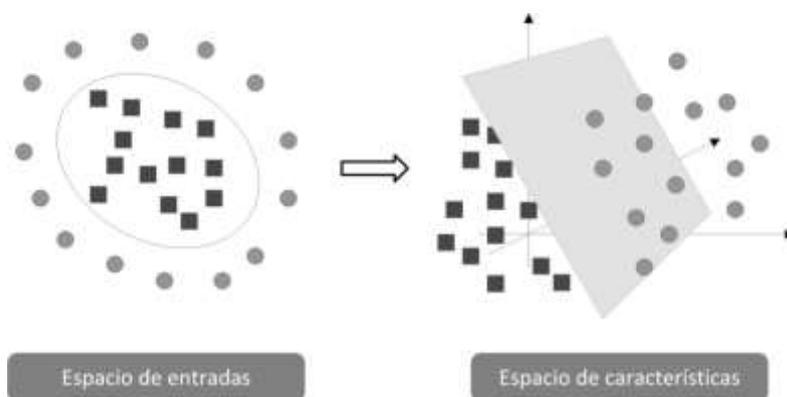


Fuente: González y Alba. (s.f.).

### 2.2.3.1.3 Support Vector Machine

Están basados de la teoría de aprendizaje estadístico, siendo su planteamiento ejecutar un método lineal para los datos en un espacio donde se tengan características de dimensión alta, relacionado de manera no lineal con el espacio de entradas. Es decir, se va a tener una transformación no lineal del espacio de entradas en uno de características de mayor dimensionalidad, donde va a ser posible poder separar de forma lineal los ejemplos de entrenamiento (véase la figura 6). La transformación se va a realizar a través de un conjunto de funciones matemáticas, las cuales se conocen como funciones núcleo (González y Alba, s.f.).

Figura 6. Transformación del espacio de entradas en un espacio de características



Fuente: González y Alba. (s.f.).

### 2.2.3.1.4 Decision tree

Un *decision tree* va a estar compuesto por un grupo de condiciones que están organizadas jerárquicamente, lo que permite que la decisión que se tome al final va a seguir todas las condiciones que se apliquen desde la raíz. Cada nodo que forma parte del árbol puede ser un nodo de decisión (una pregunta con respecto a una variable y va a distribuir una rama por cada respuesta) o un nodo hoja (decisión, predicción) como se visualiza en la figura 7 (González y Alba, s.f.).

Figura 7. Estructura de un árbol de decisión



Fuente: González y Alba (s.f.).

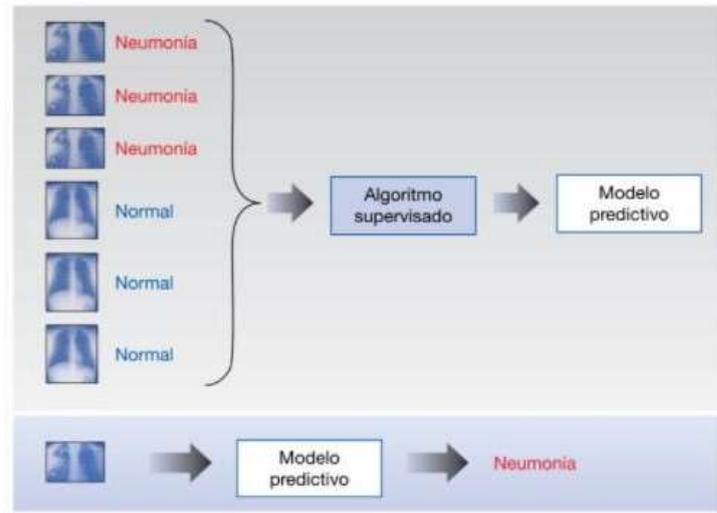
## 2.2.3.2. Tipos de aprendizaje supervisado

### 2.2.3.2.1 Aprendizaje supervisado de clasificación

Este algoritmo parte de datos etiquetados, con una variable dependiente conocida, la cual es categórica (normalmente dicotómica). Una vez se tenga entrenado el modelo predictivo con todos los datos etiquetados según evaluación, podemos realizar el modelo predictivo a los nuevos datos de los cuales no se conoce su etiquetado, por lo cual el modelo va a efectuar una predicción. Un ejemplo es realizar un algoritmo que predice la mortalidad de los pacientes que están en UCI, en donde la variable que se va a medir es “fallecido”, los cuales se clasifican entre las personas que han fallecido y las que no. Al entrenar el algoritmo se tiene información de a qué grupo forma parte cada sujeto. (Beunza, Puertas & Condés, 2019)

A continuación, en la figura 8 se presentará otro ejemplo del aprendizaje supervisado de clasificación:

Figura 8. Representación gráfica de aprendizaje supervisado de un algoritmo de clasificación



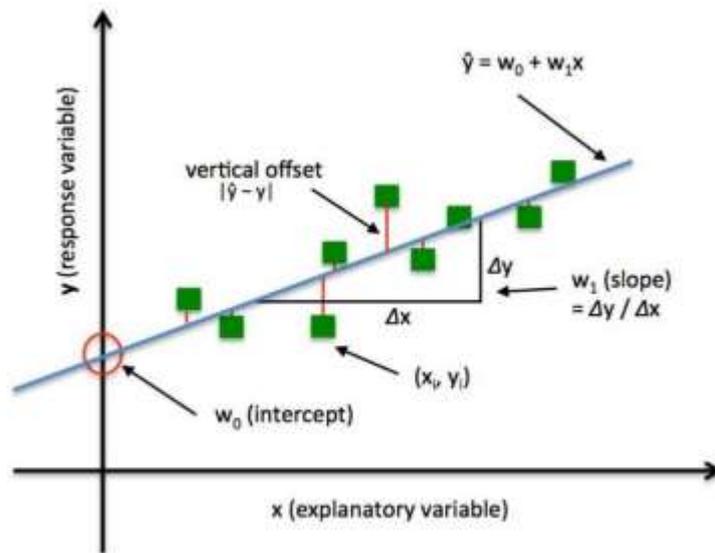
Fuente: Beunza, Puertas y Condés (2019).

#### 2.2.3.2.2 Aprendizaje supervisado de Regresión lineal

Este tipo de aprendizaje supervisado es relevante para nuestra investigación porque es el que vamos a utilizar para poder realizar el pronóstico de la demanda de la empresa en estudio. Comencemos entendiendo que es regresión lineal: Es un modelo muy atractivo por su sencilla representación, la cual “es una ecuación lineal que combina un conjunto específico de valores de entrada, cuya solución es la salida predicha para ese conjunto de valores de entrada diferente” (Barrientos & Mamani, 2019, p. 9), siendo los valores tanto de entrada como de salida numéricos.

La ecuación lineal va asignar un factor de escala a cada uno de los valores. También se va añadir un coeficiente adicional, el cual se le conoce como intersección, cuando se tiene más de una entrada ( $x$ ), la línea se va a llamar plano o hiperplano. Por lo cual su representación va a ser la forma de la ecuación y los valores usados para los coeficientes (Barrientos & Mamani, 2019). A continuación, en la figura 9 se va a representar gráficamente el modelo de aprendizaje supervisado de regresión lineal.

Figura 9. Representación gráfica del aprendizaje supervisado de regresión lineal



Fuente: Arteaga, F (2021).

Finalmente, Harrington (2012) formula la pregunta ¿Cómo se puede pasar de varios datos a la ecuación de regresión? Los datos de entrada se encuentran en la matriz  $X$  y los pesos de regresión en el vector  $w$ , para un dato  $X_I$  dado, el valor predicho sería equivalente a  $y_I = X_I^T w$ . Vamos a tener las  $X$  y las  $y$ , pero ¿Cómo encontramos las  $w$ ? Una manera es encontrar los  $w$  que minimizan el error, para lo cual vamos a definir error como la diferencia que hay entre la  $y$  predicha y la  $y$  actual. Si usamos solo el error va hacer que los valores negativos y positivos se cancelen, por lo cual se usa el error al cuadrado. Se puede escribir lo descrito como:

$$\sum_{i=1}^m (y_i - x_i^T w)^2$$

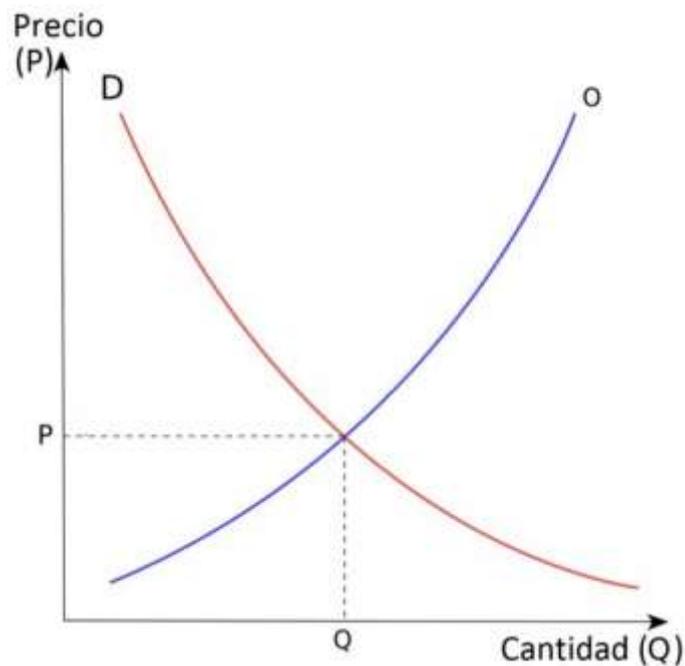
También se puede escribir en notación matricial como  $(y - Xw)^T (y - Xw)$ . Si sacamos la derivada de lo anterior con respecto a  $w$ , se obtendrá  $X^T (y - Xw)$ ; Se puede colocar dicha ecuación en cero y resolver  $w$  para tener la ecuación que a continuación la presentaremos:

$$\hat{w} = (X^T X)^{-1} X^T y$$

#### 2.2.4. Demanda

Se define como la cantidad total de un bien o servicio que la gente desea obtener. Estos bienes o servicios pueden abarcar la alimentación, educación, medicamentos, etc. En otras palabras, es todo lo que la gente desea adquirir, lo contrario a demanda es la oferta que es todo lo que los productores están dispuestos a vender. (Peiró, 2015)

Figura 10. Curva de la oferta y demanda



Fuente: Pedrosa (2016).

#### 2.2.5. Pronóstico

Es una herramienta relevante para el manejo apropiado de la CS, la cual se utiliza para predecir la demanda y con los resultados poder tener un mejor control de la incertidumbre con respecto al costo de MP, tendencia de los precios y la planeación con respecto a la producción. (Luján, 2017)

## **CAPÍTULO III: ENTORNO EMPRESARIAL**

En el presente capítulo detallaremos la información de la empresa en estudio, su estructura organizacional, su cadena de suministros, su misión, visión, valores y objetivos, así como también sus factores internos y externos (FODA), su modelo actual de negocio y su mapa de procesos actual. Con toda esta información se busca conocer un poco más a la empresa y entender los problemas, para poder ayudar en los procesos de mejora.

### **3.1. Descripción de la empresa**

#### **3.1.1. Reseña histórica y actividad económica**

Brusko es una empresa que se dedica a la comercialización de prendas de vestir e inició sus actividades en el año 1998 bajo la estructura de negocio familiar que sigue manteniéndose hasta la actualidad. En sus inicios la empresa surgió para obtener los beneficios legales y financieros que podía obtener gracias a la formalización.

La empresa, desde sus inicios funcionó bajo la guía del gerente general y dueño quien con su experiencia en el sector la ha manejado a base de juicio de expertos, recomendaciones de personas con experiencia y a prueba y error, mientras sacaba adelante su emprendimiento.

Bajo las directrices iniciales, el negocio llegó a tener 18 tiendas y estar presente en 5 establecimientos como concesionario a nivel nacional; sin embargo, la multifuncionalidad del gerente general, la conglomeración de responsabilidades y pocas personas no daba abasto para dirigir una empresa que se expandía constantemente y poco a poco se perdía en control.

Por otro lado, el éxito rotundo de los malls y las importaciones de prendas chinas en las principales provincias peruanas hizo que los niveles de control en los procesos de la empresa decayeran y originaron en Brusko una disminución en las ventas, manejo de precios, control de inventarios y otras complicaciones que afectan el desarrollo de empresa y provocó que sea necesario reducir el número de puntos de ventas.

La empresa decidió reducir la cantidad de puntos de venta para mejorar el control y enfocarse en ser más eficientes en el rubro, quedando con 14 tiendas y 2 concesionarias. Los elementos importantes de este cambio fueron: desarrollo de los mercados C y D, desinversión en los sectores más afectados, implementación de un sistema computarizado de ventas y control de stock para todas las tiendas junto con mejoras menores.

Actualmente Brusko cuenta con puntos de venta en varios departamentos del país, actualmente en Loreto, Lambayeque, Lima, Cusco, Madre de Dios y Arequipa, junto con marcas representativas que ha desarrollado desde sus inicios y Ecomoda, marca propia de la empresa, para seguir dedicándose a la venta de prendas de vestir al por menor.

### 3.1.2. Descripción de la organización

#### 3.1.2.1. Organigrama

A continuación, en la figura 11 se presenta el organigrama de la empresa Brusko S.A.C el cual cuenta con 58 empleados y está conformada por 6 áreas claves: administración, gerencia comercial, recursos humanos, tecnología de información, logística y contabilidad.

Figura 11. Organigrama



Fuente: Elaboración propia (2021).

#### 3.1.2.2. Cadena de suministros

A continuación, en la figura 12 se representa la cadena de suministro de la empresa Brusko S.A.C, mencionando y describiendo cada uno de los agentes que intervienen.

Figura 12. Cadena de suministro



Fuente: Elaboración propia (2021).

## AGENTES EXTERNOS

Son los agentes que no se involucran directamente en la cadena de suministro, pero tienen relevancia en su desenvolvimiento, se han denotado 4 como los más relevantes: El gobierno peruano y sus regulaciones, intermediarios en las compras de productos, entidades financieras de cada agente y proveedores logísticos encargados del traslado de materiales.

## AGENTES DE LA CADENA DE SUMINISTRO

Son aquellos agentes que están involucrados directamente en la cadena de suministro de Brusko, que es el agente principal.

La empresa tiene como función principal la venta al cliente final; por lo que la mayoría de los agentes se encuentran en la parte de la demanda.

### **Agentes del lado de la demanda**

- Proveedores de insumos textiles: Son los que suministran los productos primarios o procesados necesarios para la elaboración de productos textiles como: algodón, hilos, tintes, etc.
- Fábrica textil: Son los que suministran productos acabados o básicos que van a ser comercializados como polos, pantalones, casacas, poleras, shorts, etc.
- Distribuidor: Es el agente que hace compras anticipadas a las fábricas o agrega valor a productos sin marca para proporcionarlos según pedido a los retailers; pueden entregar productos como: pantalones de la marca Brusko, polos de niño con licencia, etc.

### **Agente central**

- Corporación Brusko: Es el agente principal, se encarga de la comercialización de los productos al cliente final, cuenta con distintas marcas, tallas y modelos para satisfacer la demanda. Comercializa prendas para caballero, dama, niños y otros.

### **Agentes del lado de la oferta**

- Usuario final: Son los consumidores de prendas de vestir que hacen compras minoristas en los puntos de venta.

## **FLUJOS EN LA CADENA DE SUMINISTRO**

- Flujo de productos: Es el movimiento de productos dentro de la cadena de suministro, en el caso de Brusko va de izquierda a derecha casi en su totalidad.
- Flujo de efectivo: Es el movimiento de efectivo, letras de pago, pagaré y otros relacionados en la cadena de suministro. En el caso de Brusko va de derecha a izquierda.
- Flujo de información: Es la compartición de información entre los agentes de la cadena de suministro. En este caso fluye en ambos sentidos.

### **3.1.3. Datos generales estratégicos de la empresa**

#### **3.1.3.1. Visión, misión y valores o principios**

**Visión:** Ser la empresa líder en ventas de prendas de vestir del mercado peruano para el 2025.

**Misión:** Somos una empresa comprometida con la satisfacción de compra de nuestros clientes y generación de un mejor ambiente laboral.

## **Valores organizacionales**

### **Integridad**

Actuar con respeto y honestidad. Coherente entre lo digno y lo que digo y hago. Actuar de manera correcta y respetar los ideales de aquellos que nos rodean.

### **Responsabilidad**

Asumir las consecuencias de nuestros actos y cumplir con nuestro compromisos y obligaciones ante los demás. Ser consciente de las implicancias de nuestros actos sobre nuestro entorno.

### **Respeto**

Capacidad de reconocer y apreciar a nuestros semejantes teniendo en cuenta su valor como persona. Escuchar, atender y estar atento a nuestro entorno y toda forma de vida diferente con el fin de apoyar y ser solidario en nuestro grupo social.

### **Confianza**

Mantener un pensamiento positivo y de seguridad de las acciones y forma de pensar por parte nuestros compañeros, familiares y nosotros mismos. Mantener una conducta favorable hacia quienes nos rodean.

#### **3.1.3.2. Objetivos estratégicos**

- Incrementar el margen operativo en el 10%
- Desarrollar y comercializar marcas propias
- Capacitación continua en ventas a la fuerza comercial
- Optimizar la cadena de abastecimiento

#### **3.1.3.3. Evaluación interna y externa. FODA cuantitativo**

A continuación, en la tabla 3 se presenta la matriz FODA de la empresa Brusko S.A.C:

Tabla 3. Matriz FODA

	<b>Fortalezas</b>	<b>Debilidades</b>
<b>Factores Internos</b>	F1. Posicionamiento nacional F2. Suscritos a la ley REMYPE F3. Posesión de marcas propias F4. Precios competitivos F5. Sistema de venta propio F6. Flujo de caja constante F7. Establecimiento de marca comercial F8. Buen historial crediticio	D1. Falta de automatización para las adquisiciones. D2. Ventas sujetas a estacionalidad D3. No se cuenta con venta digital D4. Poca presencia en redes D5. Falta de personal capacitado D6. Desconocimiento de la rotación de productos en tienda.
	<b>Oportunidades</b>	<b>Amenazas</b>
<b>Factores Externos</b>	O1. Crecimiento en sector retail O2. Mayor importancia de los clientes por su imagen O3. Revolución 4.0 O4. Baja de precios en locales comerciales O5. Retiro de competidores en mercados donde se encuentra presente. O6. Incremento de las ventas en línea.	A1. Nuevos competidores A2. Ambiente político inestable en el país A3. Alianza de proveedores con competidores A4. Falta de materia prima nacional A5. Nuevo brote de pandemia

Fuente: Elaboración propia (2021).

Para el análisis de la matriz FODA cuantitativo (tabla 4), se van a tomar los 5 factores más importantes para la empresa y se va a utilizar la escala de Likert con puntuaciones que van del 1 al 5 según cada cuadrante:

**Fortalezas vs Oportunidades:** 1 cuando la fortaleza no aprovecha la oportunidad y 5 cuando la fortaleza aprovecha mejor la oportunidad.

**Fortalezas y Amenazas:** 1 cuando la fortaleza no nos permite enfrentar la amenaza y 5 cuando la fortaleza nos permite enfrentar la amenaza.

**Debilidades vs Oportunidades:** 1 cuando la debilidad no nos afecta en aprovechar la oportunidad y 5 cuando la debilidad nos afecta para aprovechar la oportunidad.

**Debilidades vs Amenazas:** 1 cuando la debilidad no activará la amenaza y 5 cuando la debilidad activará la demanda.

Tabla 4. Matriz FODA cuantitativo

	O1	O3	O4	O5	O6	Resultados	A1	A2	A3	A4	A5	Resultados
<b>F1</b>	5	3	2	5	5	4	3	1	3	1	2	2
<b>F3</b>	3	1	1	3	1	1,8	3	1	1	1	1	1,4
<b>F4</b>	5	3	1	5	4	3,6	5	2	4	1	3	3
<b>F5</b>	3	4	1	1	2	2,2	1	1	1	1	1	1
<b>F8</b>	2	2	1	3	1	1,8	3	2	2	1	2	2
Resultados	3.6	2.6	1.2	3.4	2.6		3	2.4	2.2	1	1.6	
<b>D1</b>	5	5	3	3	4	4	1	1	1	1	1	1
<b>D2</b>	2	2	3	1	4	2,4	5	4	1	5	4	3,8
<b>D3</b>	5	4	1	3	5	3,6	5	1	1	1	4	2,4
<b>D5</b>	3	5	1	1	4	2,8	5	1	1	1	1	1,8
<b>D6</b>	5	4	1	4	3	3,4	4	1	5	1	1	2,4
Resultados	4	4	1.8	2.4	4		4	1.6	1.8	1.8	2.2	

Fuente: Elaboración propia (2021).

Analizando el cuadrante de fortalezas - oportunidades, para el grupo de oportunidades descrita en la matriz, la fortaleza 1 (posicionamiento nacional) es la más relevante para poder aprovechar las oportunidades y para el grupo de fortalezas, la oportunidad más probable de desarrollar es la oportunidad 1 (crecimiento), por el contrario, no se tiene las fortalezas que les permita tomar ventaja de la oportunidad 4 (precios bajos de nuevos locales).

Analizando el cuadrante de debilidades - oportunidades, nos dice que tanto la falta de automatización para adquisiciones como el desconocimiento de la rotación de productos en

tienda no permite a la empresa aprovechar al máximo la oportunidad del crecimiento en el sector retail, el uso de la tecnología que trae la revolución 4.0 y las ventas e-commerce.

Analizando el cuadrante de fortalezas - amenazas, nos indica que la fortaleza de tener precios competitivos es la más trascendental que nos permite afrontar las posibles amenazas; por otro lado, la amenaza más relevante es la de nuevos competidores, para la cual siempre debemos tener estrategias para reforzar nuestras fortalezas.

Analizando el cuadrante de debilidades - amenazas, no se puede especificar con exactitud si una debilidad o varias de ellas activen alguna amenaza, pero hay que poner énfasis en la amenaza de nuevos competidores, ya que, pueden aprovechar las debilidades que tiene la empresa.

### **3.2. Modelo de negocio actual (CANVAS)**

Propuesta de valor: Brindamos gran variedad de prendas de vestir a un menor precio que en tiendas por departamento. Se cuenta con marcas nacionales e internacionales y diseños de moda. El cliente tiene la posibilidad de devolver o cambiar el producto en caso no sea de su agrado en un plazo establecido. El personal brinda su ayuda para que pueda encontrar de manera rápida la prenda que desea.

Segmento de clientes: La venta se realiza al por menor en lugares con alto tráfico de personas y en zonas de desarrollo. Se centra principalmente en satisfacer la necesidad de prendas de vestir de calidad y moda para los sectores B, C y D.

Canal de distribución: Los canales de venta se realizan mediante los puestos de venta que se encuentran a nivel nacional. Se hacen publicaciones en medios locales para hacer publicidad. La entrega del producto se hace al momento de la compra (presencial).

Relación con los clientes: Esperamos siempre que el cliente tenga una buena experiencia al momento de la compra; para ello hay siempre personal de seguridad y custodia al ingresar a los puestos de venta, de la mano del protocolo de bioseguridad ante el covid. Luego, es atendido por los vendedores que lo ayudan a escoger las prendas que busquen y al final pasan por en los cajeros.

Fuente de ingresos: Las fuentes de ingreso se dan mediante la venta de prendas de vestir al por menor en los locales de venta y en pocas ocasiones en venta al por mayor desde el almacén. Los clientes pueden pagar con efectivo o mediante POS o carteras digitales.

Recursos clave: Nuestro principal recurso clave son las prendas de vestir que se distinguen en categorías: damas, caballeros, jóvenes y niños. El segundo recurso clave son los colaboradores con los que contamos pues la gran mayoría tiene contacto directo con los clientes e influyen en el nivel de ventas. El tercer recurso clave son las marcas con las que trabajamos pueden influenciar en la opción de compra.

Actividades clave: Las principales actividades que se realizan son las de compra, almacenamiento, distribución y venta. Estas dos últimas tienen mayor prioridad en comparación con las otras para la empresa.

Alianzas clave: Las alianzas de la empresa están compuestas en primer lugar por los proveedores de mercadería, con los que llevan varios años trabajando con la empresa se cuenta con línea de crédito o pagos en letras; en segundo lugar, los proveedores de servicio e insumos, los cuales nos brindan servicios y productos menores para desarrollar las actividades clave, y por último, los intermediarios comerciales o “brokers” los cuales buscan entre su cartera de proveedores productos que se ajusten a las necesidades de la empresa.

Estructura de costos: Dentro de los costos en los que incurre la empresa están los costos de financiamiento, costos administrativos y costos de operación. Los costos de operación cubren los gastos de adquisición de insumos, servicios y en gran medida los productos para la venta. Las prendas de vestir que se adquieren se dan en la modalidad de compra al por mayor, compras a confeccionistas y compra de lotes de remate. Esto hace que los precios de compra se encuentren a un nivel bajo y mantener el precio de competitividad en el mercado.

A continuación, en la figura 13 se presenta el modelo canvas con los 9 elementos analizados con respecto a la empresa Brusko S.A.:

Figura 13. Canvas

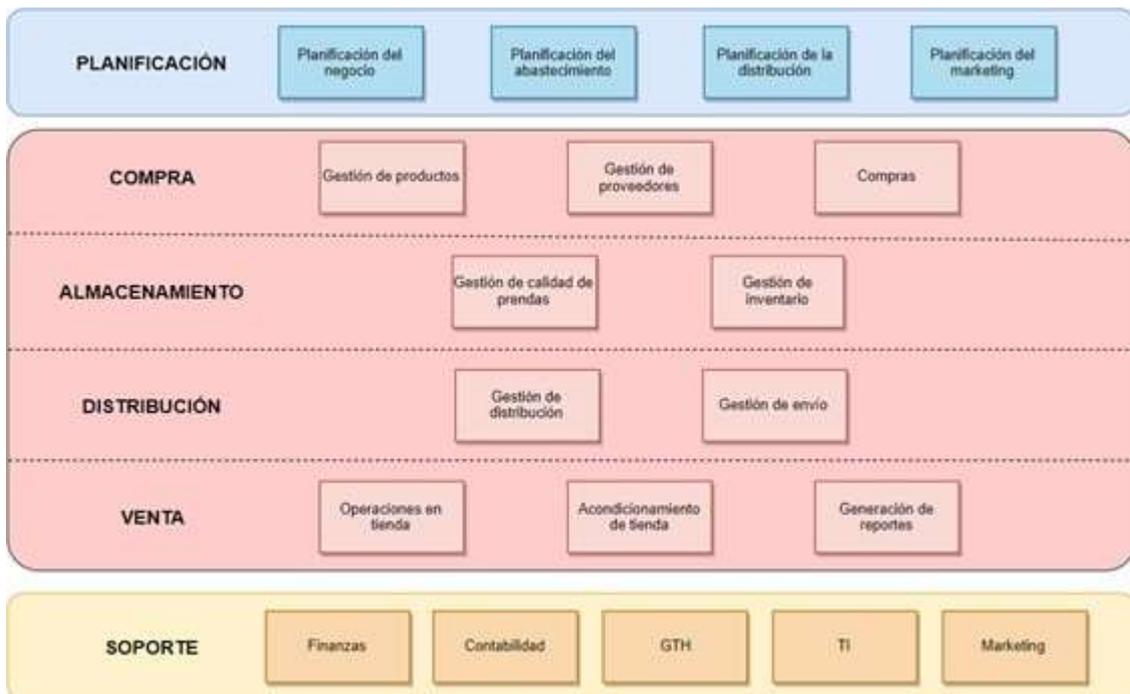
<b>CANVAS</b>				
<p><b>Socios claves</b></p> <ul style="list-style-type: none"> <li>· Proveedores de mercadería (años de relación comercial) que otorgan línea de crédito o pagos en letras.</li> <li>· Proveedores de servicio e insumos.</li> <li>· Intermediarios comerciales.</li> </ul>	<p><b>Actividades</b></p> <ul style="list-style-type: none"> <li>· Compra - almacenamiento - distribución y venta (Core).</li> </ul>	<p><b>Propuesta de valor</b></p> <ul style="list-style-type: none"> <li>· Variedad de prendas de vestir a menor precio que tiendas por departamento.</li> <li>· Marcas nacionales e internacionales.</li> <li>· Devolución o cambio del producto en un plazo establecido.</li> <li>· Personal al servicio del cliente.</li> <li>· Distintos medios de pago.</li> </ul>	<p><b>Relación con el cliente</b></p> <ul style="list-style-type: none"> <li>· Recibimiento al ingresar al punto de venta.</li> <li>· Atención por parte de vendedores. Compra rápida en caja.</li> </ul>	<p><b>Segmento de clientes</b></p> <ul style="list-style-type: none"> <li>· Personas de los sectores B, C y D</li> <li>· La venta se realiza al por menor en lugares concurridos y zonas de desarrollo</li> </ul>
	<p><b>Recursos clave</b></p> <ul style="list-style-type: none"> <li>· Las prendas de vestir</li> <li>· Los colaboradores.</li> <li>· Las marcas nacionales e internacionales.</li> </ul>		<p><b>Canales de venta</b></p> <ul style="list-style-type: none"> <li>· Puestos de venta a nivel nacional.</li> <li>· Publicaciones en medios locales</li> <li>· La entrega del producto de manera inmediata.</li> </ul>	
<p><b>Estructura de costos</b></p> <ul style="list-style-type: none"> <li>· Costos de financiamiento, costos administrativos.</li> <li>· Costos de operación: gastos de adquisición de insumos, servicios y en una gran medida los productos para la venta.</li> <li>· Las prendas de vestir se compran al por mayor, compras a confeccionistas y compra de lotes de remate. Manteniendo un precio competitivo en el mercado</li> </ul>			<p><b>Fuentes de ingreso</b></p> <ul style="list-style-type: none"> <li>· Venta de prendas de vestir al por menor en los puntos de venta y en pocas ocasiones en venta al por mayor desde el almacén.</li> <li>· Pago mediante efectivo, tarjeta y cartera virtual.</li> </ul>	

Fuente: Elaboración propia (2021).

### 3.3. Mapa de procesos actual

El mapa de procesos de la empresa que se presenta en la figura 14, visto de manera global, cuenta con 18 procesos que están agrupados dentro de las tres partes fundamentales de un negocio: Planificación, Operaciones CORE y Operaciones de soporte.

Figura 14. Mapa de procesos de Corporación Brusko S.A.C



Fuente: Elaboración propia (2021).

## CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

En el presente capítulo identificamos el enfoque, alcance y diseño de la presente investigación. Asimismo, indicaremos la metodología a usar y el cronograma de la presente propuesta.

### 4.1. Diseño de la Investigación

#### 4.1.1. Enfoque de la investigación

El enfoque de esta investigación es cuantitativo porque se concentra en las mediciones numéricas, utiliza los análisis estadísticos para responder las preguntas de investigación. (Ortega, 2018) En este caso analizaremos las ventas de pantalones de caballeros de la empresa para calcular la demanda estimada.

#### 4.1.2. Alcance de la investigación

El alcance de esta investigación es correlacional porque tiene como propósito evaluar la relación entre dos o más variables, el pronóstico de la demanda y el algoritmo de regresión lineal, en un contexto en particular. Las correlaciones se reflejan en hipótesis, y estas se someten a pruebas. (Rusu, 2011)

#### 4.1.3. Diseño o tipo de la investigación

El diseño de esta investigación es experimental porque se maneja una o más variables de estudio, controlando el aumento o disminución de estas variables para ver su efecto en las conductas observadas (Murillo, 2011).

#### 4.1.4. Población y muestra

Tabla 5. Población y muestra de la investigación

Población	Muestra
Base de datos de prendas de vestir de la empresa Brusko S.A.C.	Base de datos de pantalones de caballero de la provincia de Cusco de la empresa Brusko S.A.C desde octubre del 2018 hasta octubre del 2021.

Fuente: Elaboración propia (2021).

## 4.2. Metodología de implementación de la solución

Para lograr los objetivos de la investigación se podrían usar diferentes metodologías para un trabajo de Machine Learning, como por ejemplo la metodología KDD, SEMMA o CRISP-DM.

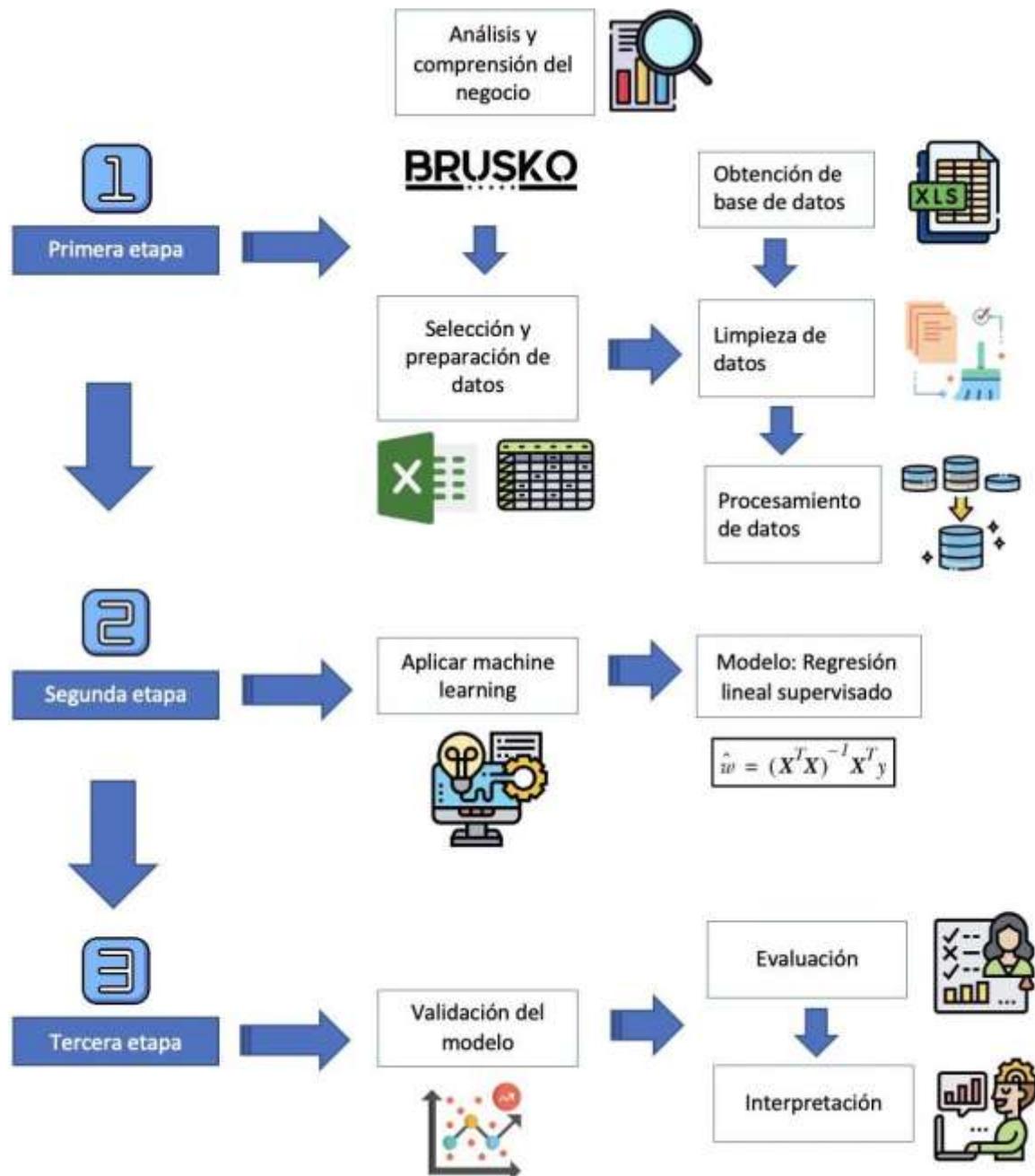
Para la presente investigación vamos a usar la metodología Crisp-DM, por ser la metodología más usada, completa y sin llegar a un nivel alto de detalle (tabla 6), ya que esta tiene un enfoque holístico, por lo que involucra al negocio para su desarrollo. Solo vamos a usar los primeros 5 pasos de la metodología Crisp-DM, ya que al ser un trabajo experimental no se procederá con el despliegue del modelo.

Tabla 6. Comparación de metodologías

FASES	KDD	CRISP – DM	SEMMA
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio	
Selección y preparación de los datos	Crear el conjunto de datos	Entendimiento de los datos	Muestreo
	Limpieza y preprocesamiento de los datos	Preparación de los datos	Comprensión
	Reducción y proyección de los datos		Modificación
Modelado	Determinar la tarea Determinar el algoritmo Modelado	Modelado	Modelado
Evaluación	Interpretación	Evaluación	Valoración
Implementación	Utilización del nuevo conocimiento	Despliegue	

Fuente: Elaboración propia (2021).

Figura 15. Diseño de la investigación



Fuente: Elaboración propia (2021).

En la primera etapa, es donde recolectamos la base de datos de la empresa, debemos entender la data y definir los objetivos del proyecto para que, a partir de esta, realicemos una limpieza, detectando anomalías, descubriendo patrones y organizándose según se necesite.

En la segunda etapa luego de tener la data organizada se aplica machine learning, en este caso utilizamos el modelo de regresión lineal supervisado para poder pronosticar la demanda a partir de las ventas.

En la tercera etapa hacemos la validación del modelo utilizado, se realizan varias pruebas para comprobar que el modelo realmente está reflejando la realidad, y finalmente se ve si el modelo utilizado va a ser útil para poder pronosticar la demanda.

### 4.3. Metodología para la medición de resultados de la implementación

Para validar el modelo se utilizarán las métricas descritas en la tabla 7, las cuales nos ayudarán a evaluar qué tan acertada es nuestra propuesta.

Tabla 7. Metodología para medición de resultados

Variables	Responsable de la Validación	Métricas
<p>Técnica de Machine Learning: Regresión Lineal Variables Independientes</p> <ul style="list-style-type: none"> <li>➤ X1 - Semana: Variable que indica el número de semana de un periodo de tiempo.</li> <li>➤ X2 - P_venta: Variable que indica el precio promedio de venta del producto para una determinada semana.</li> </ul>	Equipo de Investigación de pronóstico de la demanda	<p>Mean Absolute Error Mean Squared Error Root Mean Squared Error</p>
<p>Demanda pronosticada: Regresión Lineal Variable Dependiente</p> <ul style="list-style-type: none"> <li>➤ Y - Cantidad_v: Variable que indica la cantidad estimada a vender del producto en una determinada semana.</li> </ul>	Analista de Ventas de la empresa Brusko	<p>Demanda proyectada según la técnica aplicada para el año 2022.</p>

Fuente: Elaboración propia (2021).

#### 4.4. Cronograma de actividades y presupuesto

A continuación, en la tabla 8 presentamos el cronograma de actividades de nuestra investigación, que se desarrolla desde la primera semana de octubre de 2021.

Tabla 8. Cronograma de actividades

Actividades	Octubre				Noviembre				Diciembre			
	Sem1	Sem2	Sem3	Sem4	Sem1	Sem2	Sem3	Sem4	Sem1	Sem2	Sem3	Sem4
<b>INICIACION</b>												
Búsqueda de la problemática	■											
Búsqueda de investigación similares	■											
<b>DESARROLLO</b>												
Recolectar base de datos		■										
Tratamiento de data			■	■								
Modelamiento				■	■							
<b>CONTROL</b>												
Validación modelo						■	■					
Resultados de la solución								■	■	■		
<b>CIERRE</b>												
Implementación de la solución											■	
Verificar resultados												■

Actividad cumplida  
 Actividad por cumplir

Fuente: Elaboración propia (2021).

Asimismo, en la tabla 9 se presenta el presupuesto referencial de la implementación de la propuesta de estimación de la demanda para la Corporación Brusko tomando en consideración nuestros honorarios.

Tabla 9. Presupuesto

Actividades	Monto	Semanas	Subtotal	Total
	<b>INICIACION</b>			
Busqueda de la problemática	S/500.00	1	S/500.00	
Busqueda de investigaciones similares	S/500.00	1	S/500.00	
<b>DESARROLLO</b>				S/2,500.00
Recolectar base de datos	S/500.00	1	S/500.00	
Tratamiento de data	S/500.00	2	S/1,000.00	
Modelamiento	S/500.00	2	S/1,000.00	
<b>CONTROL</b>				S/1,500.00
Validacion modelo	S/500.00	2	S/1,000.00	
Resultados de la solución	S/500.00	1	S/500.00	
<b>CIERRE</b>				S/2,500.00
Implementación de la solución	S/500.00	3	S/1,500.00	
Verificar resultados	S/500.00	2	S/1,000.00	
				<b>S/7,500.00</b>

Fuente: Elaboración propia (2021).

## CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

### 5.1. Propuesta solución.

El presente trabajo tiene como objetivo el desarrollo de un modelo de predicción que pueda determinar la demanda de un producto determinado para las próximas líneas de tiempo (días, semanas o meses), empleando data histórica brindada por la compañía en estudio. La solución se desarrollará usando el lenguaje de programación Python, ya que es uno de los lenguajes más usados para modelos de machine learning y cuenta con librerías que facilitan el uso de algoritmos complejos, además de ser un lenguaje de fácil aprendizaje y uso.

#### 5.1.1. Planteamiento y descripción de Actividades

Para llevar a cabo las actividades en el presente trabajo, se usó la metodología definida en la figura 15, la cual está dividida en tres etapas, pero que a su vez consta de las cinco primeras fases de la metodología Crisp-DM.

**Etapa 1:** Se detallan las tres primeras fases de la metodología Crisp-DM.

Se comenzó con la **Fase 1** - Comprensión del negocio: El cual se realizó todo lo necesario para comprender al negocio y conocer sus objetivos propuestos, como se muestra en el capítulo III del presente trabajo.

Luego de ello se realizó la **Fase 2** - Comprensión de los datos: En donde recopilamos los datos brindados por la organización, entendemos los datos, exploramos y verificamos la calidad de los datos.

Seguido por la **Fase 3** - Preparación de los datos: en donde seleccionamos los datos a usar, realizamos la limpieza y construimos el dataset final.

**Etapa 2:** Se detalla la fase cuatro de la metodología Crisp-DM.

Luego pasamos a la **Fase 4** - Modelado: Donde seleccionamos las técnicas a usar, diseñamos, construimos y evaluamos el modelo.

**Etapa 3:** Se detalla la fase cinco de la metodología Crisp-DM.

Por último, pasaremos a la **Fase 5** - Evaluación de resultados: Donde evaluaremos los resultados para determinar qué tan útil es el modelo y algoritmo propuesto para la toma de decisiones en la organización.

### 5.1.2. Desarrollo de actividades. Aplicación de herramientas de solución.

#### Fase 1: Comprensión del negocio

Para la correcta comprensión del negocio, se elaboró la recopilación del entorno empresarial de la compañía Brusko (detallada en el Capítulo III). Se empezó por conocer la empresa: identificar los hechos más relevantes de la compañía a través de su desarrollo histórico, describir la estructura de trabajo, visión, misión, datos estratégicos, modelo de negocio, cadena de suministro y mapa de procesos.

#### Fase 2: Comprensión de los datos

Los datos fueron proporcionados por la Corporación Brusko S.A.C. con el objetivo de lograr el desarrollo del modelo de pronóstico de la demanda. En una primera instancia se nos mostró la estructura de la base de datos que recolecta el sistema de la compañía mediante reportes de inventario en un establecimiento de venta en donde figuran todos sus productos.

Figura 16. Estructura de la base de datos de la compañía

II	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
Fecha	Código	OP	Talla	Marca	Modelo	Temporax	Material	Descripción	Stk I	Ingre	Salid	Dev	Otra	Stoc	Prec	Stoc	PVEN	TIPO	PREN
18-01-2020	N487		4976	U	YOSHIRO			PANT DRILL NIÑO YOSHIRO	0	455	450	1	1	5	16	80		35	PANTALON
17-10-2018	N487		4623	U	YOSHIRO			PANT NIÑAS YOSHIRO	0	66	66	0	3	0	15	0		35	PANTALON
	N488		4026	U				CAMISA NIÑO A CUADROS BIG SHO	0	0	0	0	0	0	6.5	0		10	CAMISAS
05-10-2018	N489		4064	U				NIKER JEAN BENGALA NIÑAS	0	66	66	0	0	0	16	0		30	NIKER
	N489		4053	U				PANT SRTA BENGALA	0	0	0	0	0	0	17	0		35	PANTALON
	N489		4052	U				PANT. NIÑAS BENGALA	0	0	0	0	0	0	16	0		35	OVEROL
21-08-2019	N489		5070	U	SIN MARCA			PANTALONETA FRENCH NIÑA	0	120	120	0	0	0	5.3	0		10	PANTALON
09-08-2019	N490		5053	U	LEXIA			PANT JEANS NIÑA LEXIA	0	42	42	0	0	0	16	0		35	PANTALON
18-01-2020	N490		5038	U	LEXIA			PANTALON JEANS NIÑAS	0	172	172	0	1	0	16	0		35	PANTALON
	N492		4227	U				BUZO CONJUNTO NIÑOS	0	0	0	0	0	0	6	0		10	CONJUNTC
05-10-2018	N492		4226	U				POLERA CONJUNTO NIÑOS	0	4	4	0	1	0	6	0		15	CONJUNTC
05-10-2018	N493		4243	U				PANT JEAN NIÑOS ZERO	0	18	18	0	0	0	20	0		35	PANTALON
05-10-2018	N496		4474	U				LEGGINS NIÑAS COLLECTIONS	0	153	153	0	4	0	3.5	0		8	PANTALON
05-10-2018	N496		4373	U				PANTALONETA NIÑAS GIRLS	0	41	41	0	3	0	5	0		10	PANTALON
05-10-2018	N497		4543	U				BUZO CONJUNTO NIÑOS	0	46	46	0	0	0	5	0		10	BUZOS
05-10-2018	N497		4559	U				POLERA FRANELA NIÑAS	0	113	113	0	1	0	10	0		25	POLERA
05-10-2018	N497		4507	U				CASACA CONJUNTO NIÑA	0	16	16	0	0	0	10	0		20	CONJUNTC
	N497		4401	U				CASACA CNITO NIÑAS	0	0	0	0	0	0	10	0		20	CONJUNTC
	N497		4402	U				BUZO CNITO NIÑAS	0	0	0	0	0	0	5	0		10	CONJUNTC
18-07-2019	N497		5027	U	SIN MARCA			BUZO FRANELA NIÑO	0	155	155	0	0	0	6.7	0		10	BUZOS
05-10-2018	N497		4508	U				BUZO CONJUNTO NIÑA	0	13	13	0	0	0	5	0		10	CONJUNTC
05-10-2018	N497		4542	U				POLERA CONJUNTO NIÑOS	0	45	45	0	1	0	10	0		20	POLERA
05-10-2018	N497		4381	U				BUZO CNITO NIÑOS	0	9	9	0	0	0	5	0		10	PANTALON
11-06-2019	N497		4992	U	SIN MARCA			BUZO FRANELA NIÑO	0	610	610	0	5	0	6.6	0		10	BUZOS
29-09-2020	N497		5658	U	KIDS			POLERA FLOREADA NIÑA KIDS FOX	0	134	117	0	1	17	11	187		25	POLERA

Fuente: Corporación Brusko S.A.C. (2021).

Debido a la variabilidad de tipos de prendas se decidió tomar solo un tipo de producto; por recomendación del especialista en la compañía, se toma los pantalones de caballero por ser

el producto que cuenta con una rotación, disponibilidad de los datos y margen de ganancia altos.

Se segmentan los datos que pudo proporcionar la compañía y de ellas se tomó la provincia de Cusco como punto de muestreo, la cual es la que mayor volumen de venta mantenía con respecto a las otras según señala el especialista.

Luego de la segmentación como producto de estudio los pantalones de caballero por ser un producto importante para la empresa; se hizo una solicitud del detalle 334 reportes de productos que coincidían con las características del producto en estudio para su análisis los cuales serán tomados como la demanda del producto.

Figura 17. Extracción de reportes de pantalones de caballero

The screenshot shows a software window titled "MOVIMIENTO POR CÓDIGO". At the top, there is a search form with the following fields: "CÓDIGO" (00005754u), "TIPO MOV" (empty), "INICIO" (25/01/2018), and "FIN" (25/01/2021). A "Buscar" button is located to the right of the date fields. Below the search form, there are two tables: "Ingresos" and "Salidas".

**Ingresos Table:**

Documento	Correlativa	Código	Descripción	Fecha	TipoMov	Cantidad
GUIA INTERNA	00028742	00005754U	PANT DRILL CAB VARIOS APACI	09-12-2020	COMPRAS DE MERCADERIA	200
REGULARIZACION	00000013	00005754U	PANT DRILL CAB VARIOS APACI	21-01-2021	INGRESOS AJUSTE DE INVENTAR	3

**Total Ingresos:** 203

**Salidas Table:**

Documento	Correlativa	Código	Descripción	Fecha	TipoMov	Cantidad
BOLETA VENTA ELECT	00002375	00005754U	PANT DRILL CAB VARIOS APACI	31-12-2020	VENTAS	1
BOLETA VENTA ELECT	00002394	00005754U	PANT DRILL CAB VARIOS APACI	31-12-2020	VENTAS	1
BOLETA VENTA ELECT	00002411	00005754U	PANT DRILL CAB VARIOS APACI	31-12-2020	VENTAS	2
BOLETA VENTA ELECT	00002476	00005754U	PANT DRILL CAB VARIOS APACI	31-12-2020	VENTAS	1
BOLETA VENTA ELECT	00002501	00005754U	PANT DRILL CAB VARIOS APACI	31-12-2020	VENTAS	1
BOLETA VENTA ELECT	00002516	00005754U	PANT DRILL CAB VARIOS APACI	02-01-2021	VENTAS	1
BOLETA VENTA ELECT	00002540	00005754U	PANT DRILL CAB VARIOS APACI	02-01-2021	VENTAS	1
BOLETA VENTA ELECT	00048147	00005754U	PANT DRILL CAB VARIOS APACI	04-01-2021	VENTAS	1
BOLETA VENTA ELECT	00048176	00005754U	PANT DRILL CAB VARIOS APACI	04-01-2021	VENTAS	1
BOLETA VENTA ELECT	00048320	00005754U	PANT DRILL CAB VARIOS APACI	06-01-2021	VENTAS	1
BOLETA VENTA ELECT	00048335	00005754U	PANT DRILL CAB VARIOS APACI	06-01-2021	VENTAS	1

**Total Salidas:** 177

Fuente: Corporación Brusko S.A.C. (2021).

Se obtuvieron los datos de la demanda a través de reportes en Excel con los que trabaja la empresa. El encargado de la compañía nos brindó la información consolidada de todos los reportes de pantalones de caballero desde el año 2018 hasta el 2021 con la siguiente estructura:

Tipo de documento, Número de documento, Código único de prenda, Descripción del producto, Fecha de venta, Tipo de Movimiento, y cantidad vendida.

Tabla 10. Reporte de detalle de la demanda del producto

Documento	Correlativo	Código	Descripción	Fecha	TipoMov	Cantidad
BOLETA VENT	21137	00004065U	PANT ANALOG CABALLERO	30-06-2019	VENTAS	1
BOLETA VENT	21492	00004065U	PANT ANALOG CABALLERO	06-07-2019	VENTAS	1
BOLETA VENT	23668	00004065U	PANT ANALOG CABALLERO	01-08-2019	VENTAS	1
BOLETA VENT	4621	411328	PANT. APACHE JEAN CLASICO	08-12-2018	VENTAS	1
BOLETA VENT	2	411328	PANT. APACHE JEAN CLASICO	10-12-2018	VENTAS	1
BOLETA VENT	5523	411328	PANT. APACHE JEAN CLASICO	16-12-2018	VENTAS	1
BOLETA VENT	10095	411328	PANT. APACHE JEAN CLASICO	20-01-2019	VENTAS	2
BOLETA VENT	12503	411328	PANT. APACHE JEAN CLASICO	22-02-2019	VENTAS	1
BOLETA VENT	12537	411328	PANT. APACHE JEAN CLASICO	22-02-2019	VENTAS	1
BOLETA VENT	12577	411328	PANT. APACHE JEAN CLASICO	23-02-2019	VENTAS	1
BOLETA VENT	9065	411330	PANT. APACHE JEAN CLASICO	07-01-2019	VENTAS	1
BOLETA VENT	1269	00004116U	PANT MODA CAB BROOKLIN	23-10-2018	VENTAS	1
BOLETA VENT	2	00004116U	PANT MODA CAB BROOKLIN	10-12-2018	VENTAS	1
BOLETA VENT	10061	00004116U	PANT MODA CAB BROOKLIN	20-01-2019	VENTAS	1
BOLETA VENT	10710	00004116U	PANT MODA CAB BROOKLIN	30-01-2019	VENTAS	1
BOLETA VENT	406	416928	PANT PACHECLASICOCAB	09-10-2018	VENTAS	1
BOLETA VENT	616	00004177U	PANT CLAS-CO CAB APACHE	13-10-2018	VENTAS	1
BOLETA VENT	827	00004177U	PANT CLAS-CO CAB APACHE	15-10-2018	VENTAS	1
BOLETA VENT	3777	00004178U	PANT BROOKLING PREMIUM	26-11-2018	VENTAS	2
BOLETA VENT	8242	00004178U	PANT BROOKLING PREMIUM	31-12-2018	VENTAS	2
BOLETA VENT	7384	00004179U	PANTSTRETCH COLOMB CAB	26-12-2018	VENTAS	2
BOLETA VENT	2535	427236	PANT JEAN BRUSKO CLASICO	10-11-2018	VENTAS	1
BOLETA VENT	3181	427236	PANT JEAN BRUSKO CLASICO	18-11-2018	VENTAS	1
BOLETA VENT	4129	427236	PANT JEAN BRUSKO CLASICO	01-12-2018	VENTAS	1
BOLETA VENT	5700	427236	PANT JEAN BRUSKO CLASICO	18-12-2018	VENTAS	1
BOLETA VENT	6137	427236	PANT JEAN BRUSKO CLASICO	21-12-2018	VENTAS	1
BOLETA VENT	13060	427236	PANT JEAN BRUSKO CLASICO	02-03-2019	VENTAS	1
BOLETA VENT	4	427236	PANT JEAN BRUSKO CLASICO	09-05-2019	VENTAS	5
BOLETA VENT	19726	427236	PANT JEAN BRUSKO CLASICO	09-06-2019	VENTAS	1
BOLETA VENT	19747	427236	PANT JEAN BRUSKO CLASICO	09-06-2019	VENTAS	1

Fuente: Corporación Brusko S.A.C. (2021)

### Fase 3: Preparación de los datos

En el presente trabajo nos enfocaremos en las ventas de un solo punto de venta y en un solo tipo de producto para poder facilitar el análisis, si se requiere analizar otros productos o las ventas en otras provincias solo bastaría replicar el modelo. El producto escogido son los pantalones de caballero, por lo que de los datos obtenidos se preprocesaron.

Se arregló los datos que fueron proporcionados por la empresa de forma manual antes de llevarlo a la herramienta para la programación; en primer lugar, se consolida la demanda de unitaria a demanda semanal para poder usarlo como variable.

Mediante la estructura de datos inicial proporcionado por la empresa, se pudo obtener los precios de venta promedio del tipo de producto haciendo el cruce con el código único de

cada producto fue agregado. Este precio promedio será considerado para ser usado como variable.

Finalmente se encontraron datos atípicos que corresponden al inicio y al final de la cuarentena obligatoria que afectó al año 2020, esto fue validado por el especialista de la empresa y posteriormente se separaron del modelo para ver su resultado.

Tabla 11. Reporte de ventas semanales pantalones caballero en Cusco

Document	Correlat	Código	Descripción	Fecha	TipoMc	Semar	Semana_c	P_venta	Canti
BOLETA VENT	53	448842	PANT JEAN CI	05/10/2018	VENTAS	40	1	49.47368	20
BOLETA VENT	278	427228	PANT JEAN BF	07/10/2018	VENTAS	41	2	47.30612	56
BOLETA VENT	686	456034	PANT JEANS C	14/10/2018	VENTAS	42	3	46.77561	46
BOLETA VENT	1161	459130	PANT DRILL C	21/10/2018	VENTAS	43	4	43.85833	38
BOLETA VENT	1637	00004465U	PANT DRILL C	28/10/2018	VENTAS	44	5	47.03148	59
BOLETA VENT	2172	456032	PANT JEANS C	04/11/2018	VENTAS	45	6	47.93973	76
BOLETA VENT	2661	456034	PANT JEANS C	11/11/2018	VENTAS	46	7	47.86304	49
BOLETA VENT	3181	427236	PANT JEAN BF	18/11/2018	VENTAS	47	8	47.27385	70
BOLETA VENT	3715	456032	PANT JEANS C	25/11/2018	VENTAS	48	9	48.67385	75
BOLETA VENT	4212	448828	PANT JEAN CI	02/12/2018	VENTAS	49	10	47.84304	91
BOLETA VENT	4851	456030	PANT JEANS C	09/12/2018	VENTAS	50	11	47.12903	116
BOLETA VENT	5523	411328	PANT. APACH	16/12/2018	VENTAS	51	12	48.50413	135
BOLETA VENT	6584	459130	PANT DRILL C	23/12/2018	VENTAS	52	13	48.09824	187
BOLETA VENT	8058	441828	PANT DRILL S	30/12/2018	VENTAS	53	14	47.68022	96
BOLETA VENT	8560	427228	PANT JEAN BF	02/01/2019	VENTAS	1	15	47.14815	94
BOLETA VENT	8981	00004065U	PANT ANALOC	06/01/2019	VENTAS	2	16	46.42714	81

Fuente: Corporación Brusko S.A.C. (2021)

#### Fase 4: Modelado

Para poder realizar el modelado se usó el lenguaje de programación Python bajo la plataforma de Anaconda y usando la interfaz de Jupyter Notebook.

Para poder trabajar el modelo de predicción con la data de demanda obtenida, esta se dividió en tres partes correspondiendo a cada año de venta (2018, 2019, 2020 y 2021) y se tiene un conglomerado de todos los años. El modelo trabaja con el 80% de la muestra como data de entrenamiento y el 20% restante como data de test o prueba.

Como técnica de modelado y por el tipo de solución a ofrecer, se escogió la regresión lineal, técnica extraída de la biblioteca de scikit-learn o sklearn, biblioteca que contiene varias herramientas o algoritmos para el análisis de datos, tales como, svm, knn, random forest, k-means, entre otros.

#### Fase 5: Evaluación de resultados

El modelo propuesto fue el de regresión lineal, el cual fue programado en python mediante la interfaz de jupyter. La evaluación de estos se realizará en la sección 5.2 Medición de la solución.

## 5.2. Medición de la solución.

Para medir la solución se usaron las métricas MAE, MSE y RMSE, que son el error absoluto medio, error cuadrático medio y la raíz del error cuadrático medio respectivamente, los cuales se describirán a continuación:

**MAE**, tiene como objetivo evaluar la calidad de la técnica de machine learning, la cual se refiere a la diferencia entre el valor predicho y lo real (Gamarra, 2019).

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

- $Y_i$  es el valor pronosticado para el  $i$ ésimo observación en el conjunto de datos.
- $X_i$  es el valor real para el  $i$ ésimo observación en el conjunto de datos.
- $n$  es el tamaño de la muestra.

**MSE**, señala que tan cerca está la línea de regresión de un grupo de puntos (Yu-Sun, Chun-Hung & Shingchern, 2021).

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_{obs,i} - x_{model,i})^2$$

- $X_{obs,i}$  es el valor observado para el  $i$ ésimo observación en el conjunto de datos.
- $X_{model,i}$  es el valor pronosticado para el  $i$ ésimo observación en el conjunto de datos.
- $n$  es el tamaño de la muestra.

**RMSE**, su objetivo es medir la dispersión de los datos con respecto a la línea de regresión. Calcula la distancia promedio entre los valores observados del modelo y los valores pronosticados en el grupo de datos (Gamarra, 2019).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{obs,i} - x_{model,i})^2}$$

- $X_{obs,i}$  es el valor observado para el  $i$ ésimo observación en el conjunto de datos.

- $X_{\text{model},i}$  es el valor pronosticado para el  $i$ ésimo observación en el conjunto de datos.
- $n$  es el tamaño de la muestra.

### 5.2.1. Análisis de Indicadores cuantitativo y/o cualitativo.

Tras aplicar el modelo de regresión lineal a la data brindada por la empresa, se obtuvieron los siguientes resultados para los dos escenarios propuestos:

Tabla 12. Cuadro comparativo de métricas de regresión

<b>Dataset / Métricas</b>	<b>MAE</b>	<b>MSE</b>	<b>RMSE</b>
Dataset 2018 - 2021	33.8059	3489.0712	59.0684
Dataset 2018	19.8282	565.6315	23.7830
Dataset 2019	10.6155	174.7536	13.2194
Dataset 2020	40.3999	2220.0844	47.1178
Dataset 2021	15.1161	319.2429	17.8674
Dataset 2018 - 2021 SO	21.9213	898.5669	29.9761

Fuente: Elaboración propia (2021).

#### Mean absolute error (MAE)

De la tabla 12, podemos afirmar que el modelo con el dataset del 2019 presenta un menor MAE respecto a los demás datasets evaluados. Este valor (10.6155) nos indica que los valores predichos en promedio difieren en +/- 10.62 unidades respecto a los datos observados.

#### Mean squared error (MSE)

De los datos obtenidos, se observa que presentamos valores muy diversos (174.7536 - 3489.0712). Si se considera el dataset 2018 -2021 y dataset 2020 el modelo tendría un error significativo; sin embargo, si tomamos el modelo del dataset 2019 la varianza del error de pronóstico sería mucho menor en comparación a los demás modelos.

#### Root Mean Squared Error (RMSE)

De los 6 modelos evaluados, podemos afirmar que si se considera el modelo con el dataset 2018 - 2021 sin eliminar los outliers obtendremos un error cuadrático medio de 59.0684. Este valor, al ser muy elevado, nos indica que los valores observados no están tan cerca a los

valores predichos del modelo, a diferencia del modelo utilizando el dataset 2019 en el cual se obtiene un RMSE de 13.2194

Tras comparar los valores de todas las métricas obtenidas, se concluye que los valores predichos con el dataset 2019 encajan de mejor manera, en comparación a los otros resultados.

### 5.2.2. Simulación de solución. Aplicación de Software.

A continuación, mostraremos el modelo usado para el pronóstico de la demanda, usando la técnica de regresión lineal:

Se inició importando la librería pandas, numpy y matplotlib para poder leer nuestro dataset, para poder realizar operaciones matemáticas y para poder realizar gráficos de dispersión.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Comenzamos leyendo nuestras bases de datos por año ('datasetaño') y en su totalidad ('dataset'). El último dataset ('datasetso'), se propone tras eliminar los outliers que se presentaron en la semana de inicio de pandemia (2020).

```
dataset2018 = pd.read_excel('bd_P2018.xlsx')
dataset2019 = pd.read_excel('bd_P2019.xlsx')
dataset2020 = pd.read_excel('bd_P2020.xlsx')
dataset2021 = pd.read_excel('bd_P2021.xlsx')
dataset = pd.read_excel('bd_totalP.xlsx')
datasetso = pd.read_excel('bd_totalP_sin_outliers.xlsx')
```

Verificamos las dimensiones de nuestros datos.

```
d2018=dataset2018.shape
d2019=dataset2019.shape
d2020=dataset2020.shape
d2021=dataset2021.shape
dBD=dataset.shape
dBD50=datasetso.shape
print(d2018,d2019,d2020,d2021,dBD,dBD50)
```

(14, 9) (53, 9) (40, 9) (40, 9) (147, 10) (145, 10)

Obtenemos los tipos de datos de nuestro dataframe.

```
dataset.dtypes
```

```

Documento          object
Correlativo        int64
Código             object
Descripción        object
Fecha             datetime64[ns]
TipoMov            object
Semana             int64
Semana_cont        int64
P_venta            float64
Cantidad_v         int64
dtype: object

```

Mostramos los 10 primeros datos del dataframe 'dataset', que contiene todos los años en estudio.

```
dataset.head(10)
```

Tabla 13. Dataframe de los años 2018-2021

	Documento	Correlativo	Código	Descripción	Fecha	TipoMov	Semana	Semana_cont	P_venta	Cantidad_v
0	BOLETA VENTA ELECTRONICA	53	448842	PANT JEAN CLASICO BRUSKO CAB	2018-10-05	VENTAS	40	1	49.473684	20
1	BOLETA VENTA ELECTRONICA	278	427228	PANT JEAN BRUSKO CLASICO	2018-10-07	VENTAS	41	2	47.306122	56
2	BOLETA VENTA ELECTRONICA	686	456034	PANT JEANS CLASICO APACHE	2018-10-14	VENTAS	42	3	46.775610	46
3	BOLETA VENTA ELECTRONICA	1161	459130	PANT DRILL CLASICO APACHE	2018-10-21	VENTAS	43	4	43.858333	38
4	BOLETA VENTA ELECTRONICA	1637	00004465U	PANT DRILL OV7 CAB	2018-10-28	VENTAS	44	5	47.031481	59
5	BOLETA VENTA ELECTRONICA	2172	456032	PANT JEANS CLASICO APACHE	2018-11-04	VENTAS	45	6	47.939726	76
6	BOLETA VENTA ELECTRONICA	2661	456034	PANT JEANS CLASICO APACHE	2018-11-11	VENTAS	46	7	47.863043	49
7	BOLETA VENTA ELECTRONICA	3181	427236	PANT JEAN BRUSKO CLASICO	2018-11-18	VENTAS	47	8	47.273846	70
8	BOLETA VENTA ELECTRONICA	3715	456032	PANT JEANS CLASICO APACHE	2018-11-25	VENTAS	48	9	48.673846	75
9	BOLETA VENTA ELECTRONICA	4212	448828	PANT JEAN CLASICO BRUSKO CAB	2018-12-02	VENTAS	49	10	47.843038	91

Fuente: Elaboración propia (2021)

Mostramos los 10 primeros datos del dataframe del año 2019 como muestra.

```
dataset2019.head(10)
```

Tabla 14. Dataframe del año 2019

	Documento	Correlativo	Código	Descripción	Fecha	TipoMov	Semana	P_venta	Cantidad_v
0	BOLETA VENTA ELECTRONICA	8560	427228	PANT JEAN BRUSKO CLASICO	2019-01-02	VENTAS	1	47.148148	94
1	BOLETA VENTA ELECTRONICA	8981	00004065U	PANT ANALOG CABALLERO	2019-01-06	VENTAS	2	46.427143	81
2	BOLETA VENTA ELECTRONICA	9595	00004065U	PANT ANALOG CABALLERO	2019-01-13	VENTAS	3	47.435484	66
3	BOLETA VENTA ELECTRONICA	10095	411328	PANT. APACHE JEAN CLASICO	2019-01-20	VENTAS	4	48.080000	53
4	BOLETA VENTA ELECTRONICA	10517	448828	PANT JEAN CLASICO BRUSKO CAB	2019-01-27	VENTAS	5	46.862069	74
5	BOLETA VENTA ELECTRONICA	11030	459132	PANT DRILL CLASICO APACHE	2019-02-03	VENTAS	6	46.414286	88
6	BOLETA VENTA ELECTRONICA	11692	459128	PANT DRILL CLASICO APACHE	2019-02-10	VENTAS	7	47.765957	56
7	BOLETA VENTA ELECTRONICA	12217	481430	PANT JEAN APACHE CLASICO	2019-02-17	VENTAS	8	47.672414	73
8	BOLETA VENTA ELECTRONICA	12668	481430	PANT JEAN APACHE CLASICO	2019-02-24	VENTAS	9	46.510256	50
9	BOLETA VENTA ELECTRONICA	13178	456028	PANT JEANS CLASICO APACHE	2019-03-03	VENTAS	10	48.743243	86

Fuente: Elaboración propia (2021)

Seleccionamos las variables a usar para poder obtener los principales estadísticos por cada dataset trabajado (ver tabla 15, tabla 16, tabla 17, tabla 18, tabla 19 y tabla 20).

```
dataset[['Semana_cont', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 15. Tabla de los valores estadísticos 2018-2021

	Semana_cont	P_venta	Cantidad_v
<b>count</b>	147.000000	147.000000	147.000000
<b>mean</b>	74.000000	47.289368	81.959184
<b>std</b>	42.579338	1.523654	41.960986
<b>min</b>	1.000000	43.216216	4.000000
<b>25%</b>	37.500000	46.420714	60.000000
<b>50%</b>	74.000000	47.472222	75.000000
<b>75%</b>	110.500000	48.512353	93.000000
<b>max</b>	147.000000	50.000000	348.000000

Fuente: Elaboración propia (2021)

De la tabla 15 podemos observar que el dataset consta de 147 semanas analizadas, el precio de venta promedio por producto es de S/ 47.29 y que la cantidad promedio de prendas vendidas fueron 82 durante los años 2018-2021.

```
dataset2018[['Semana', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 16. Tabla de los valores estadísticos del año 2018

	Semana	P_venta	Cantidad_v
<b>count</b>	14.0000	14.000000	14.000000
<b>mean</b>	46.5000	47.532168	79.571429
<b>std</b>	4.1833	1.281005	43.662542
<b>min</b>	40.0000	43.858333	20.000000
<b>25%</b>	43.2500	47.165236	50.750000
<b>50%</b>	46.5000	47.761629	72.500000
<b>75%</b>	49.7500	48.058608	94.750000
<b>max</b>	53.0000	49.473684	187.000000

Fuente: Elaboración propia (2021)

De la tabla 16 podemos observar que el dataset consta de 14 semanas analizadas, el precio de venta promedio por producto es de S/ 47.53 y que la cantidad promedio de prendas vendidas fueron 79 durante el año 2018.

```
dataset2019[['Semana', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 17. Tabla de los valores estadísticos del año 2019

	Semana	P_venta	Cantidad_v
<b>count</b>	53.000000	53.000000	53.000000
<b>mean</b>	27.000000	47.816700	85.433962
<b>std</b>	15.443445	1.114749	28.657466
<b>min</b>	1.000000	45.175676	50.000000
<b>25%</b>	14.000000	47.032258	66.000000
<b>50%</b>	27.000000	47.716418	83.000000
<b>75%</b>	40.000000	48.743243	94.000000
<b>max</b>	53.000000	49.854839	216.000000

Fuente: Elaboración propia (2021)

De la tabla 17 podemos observar que el dataset consta de 53 semanas analizadas, el precio de venta promedio por producto es de S/ 47.82 y que la cantidad promedio de prendas vendidas fueron 85 durante el año 2019.

```
dataset2020[['Semana', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 18. Tabla de los valores estadísticos del año 2020

	Semana	P_venta	Cantidad_v
<b>count</b>	40.000000	40.000000	40.000000
<b>mean</b>	29.600000	46.372975	88.900000
<b>std</b>	16.883917	2.006309	60.784529
<b>min</b>	1.000000	43.216216	4.000000
<b>25%</b>	10.750000	44.854167	55.000000
<b>50%</b>	33.500000	45.868099	78.000000
<b>75%</b>	43.250000	48.321534	99.000000
<b>max</b>	53.000000	50.000000	348.000000

Fuente: Elaboración propia (2021)

De la tabla 18 podemos observar que el dataset consta de 40 semanas analizadas, el precio de venta promedio por producto es de S/ 46.37 y que la cantidad promedio de prendas vendidas fueron 89 durante el año 2020.

```
dataset2021[['Semana', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 19. Tabla de los valores estadísticos del año 2021

	Semana	P_venta	Cantidad_v
<b>count</b>	40.000000	40.000000	40.000000
<b>mean</b>	20.500000	47.422067	71.250000
<b>std</b>	11.690452	1.087766	31.316212
<b>min</b>	1.000000	44.523810	22.000000
<b>25%</b>	10.750000	46.774220	53.500000
<b>50%</b>	20.500000	47.390365	68.500000
<b>75%</b>	30.250000	48.380911	81.250000
<b>max</b>	40.000000	49.000000	187.000000

Fuente: Elaboración propia (2021)

De la tabla 19 podemos observar que el dataset consta de 40 semanas analizadas, el precio de venta promedio por producto es de S/ 47.42 y que la cantidad promedio de prendas vendidas fueron 71 durante el año 2021.

```
datasetso[['Semana_cont', 'P_venta', 'Cantidad_v']].describe()
```

Tabla 20. Tabla de los valores estadísticos de los años 2018-2021 sin valores outliers

	Semana_cont	P_venta	Cantidad_v
count	145.000000	145.000000	145.000000
mean	73.000000	47.260410	82.993103
std	42.001984	1.512122	41.302996
min	1.000000	43.216216	20.000000
25%	37.000000	46.414286	60.000000
50%	73.000000	47.439394	75.000000
75%	109.000000	48.504132	94.000000
max	145.000000	49.854839	348.000000

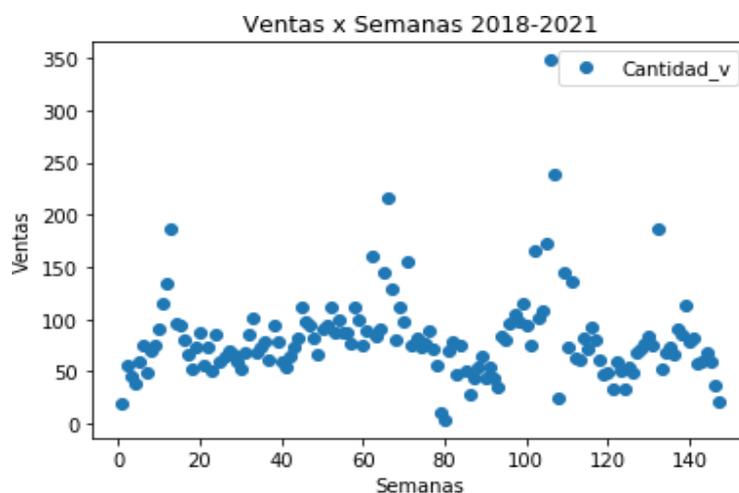
Fuente: Elaboración propia (2021)

De la tabla 20 podemos observar que el dataset varía con respecto a la tabla 15, ya que consta de 145 semanas analizadas (se eliminó 2 semanas outliers), el precio de venta promedio por producto es de S/ 47.26 y que la cantidad promedio de prendas vendidas fueron 82 durante los años 2018 -2021.

Realizamos el gráfico de dispersión entre una variable independiente (Semana) versus la dependiente (Cantidad\_V), para los años en estudio (ver figura 18, figura 19, figura 20, figura 21, figura 22, figura 23).

```
dataset.plot(x='Semana_cont', y='Cantidad_v', style="o")
plt.title('Ventas x Semanas 2018-2021')
plt.xlabel('Semanas')
plt.ylabel('Ventas')
plt.show()
```

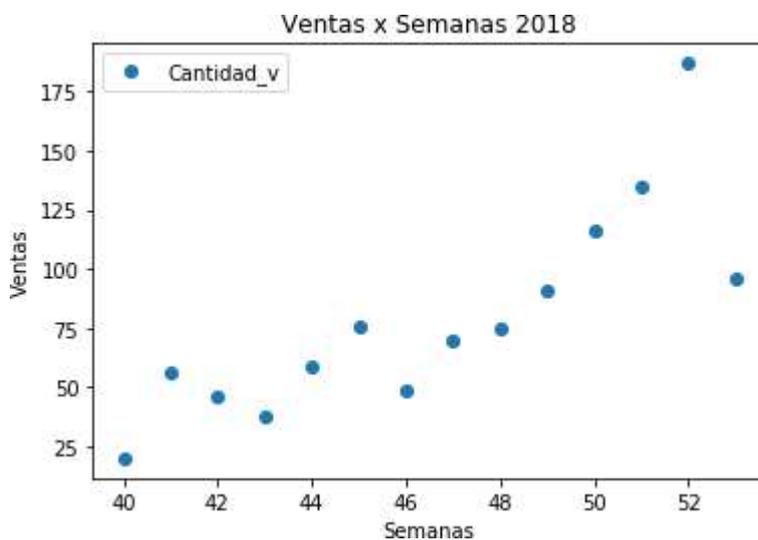
Figura 18. Gráfico de dispersión del año 2018-2021



Fuente: Elaboración propia (2021)

De la figura 18 podemos observar que cada cierta semana hay picos en las ventas, esto se debe a las ventas por campaña navideña y estas se repiten por los años en estudio.

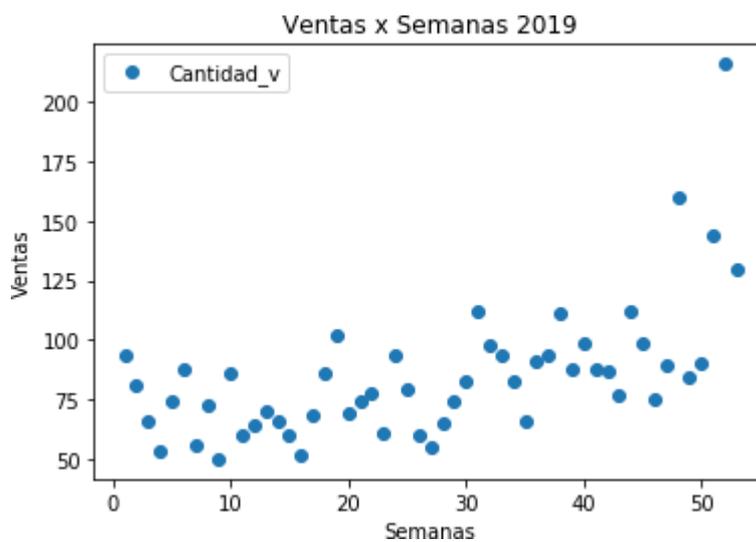
Figura 19. Gráfico de dispersión del año 2018



Fuente: Elaboración propia (2021)

De la figura 19 podemos observar que entre la semana 50 a la 52 hay picos, esto se debe a las ventas por campaña navideña.

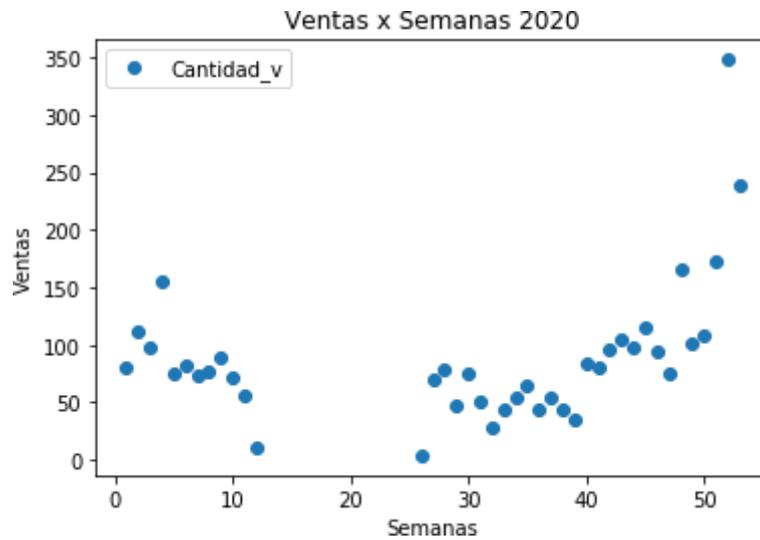
Figura 20. Gráfico de dispersión del año 2019



Fuente: Elaboración propia (2021)

De la figura 20 podemos observar que entre la semana 50 a la 52 hay picos, esto se debe a las ventas por campaña navideña.

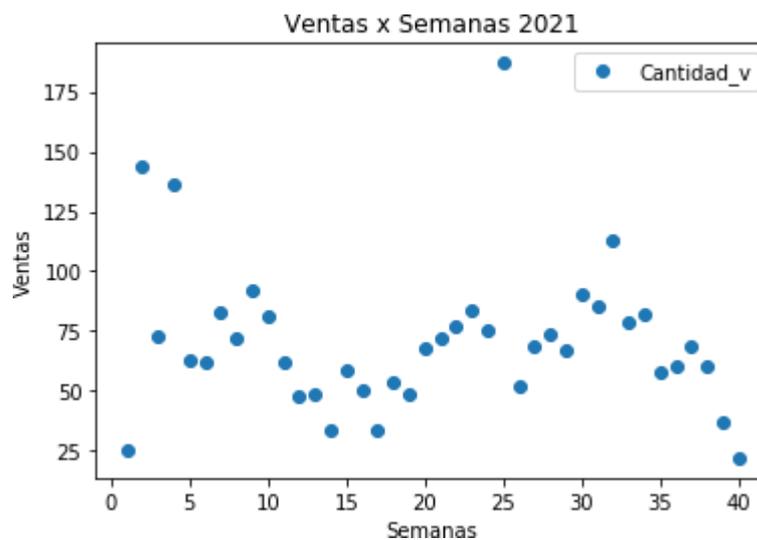
Figura 21. Gráfico de dispersión del año 2020



Fuente: Elaboración propia (2021)

De la figura 21 podemos observar que entre la semana 12 a la 24 existe un vacío, esto se debe a que en dichas semanas no se realizaron ventas por la pandemia. De igual forma podemos observar que entre la semana 50 a la 52 hay picos, esto se debe a las ventas por campaña navideña.

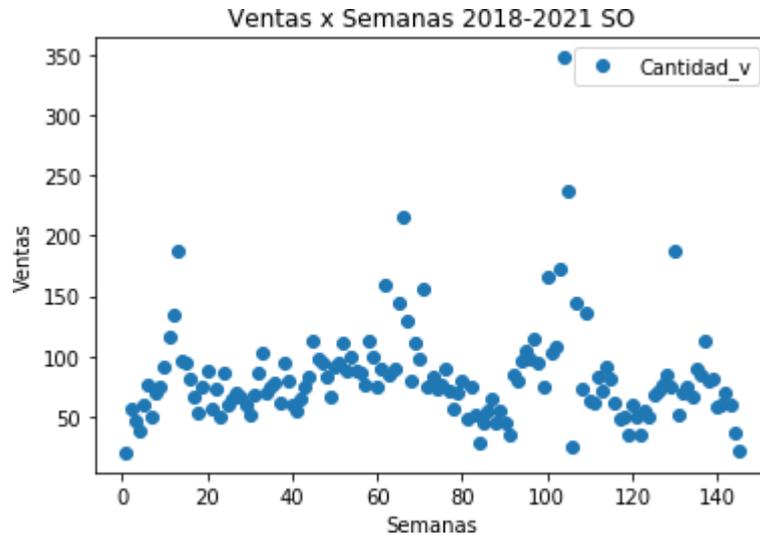
Figura 22. Gráfico de dispersión del año 2021



Fuente: Elaboración propia (2021)

De la figura 22 podemos observar que existen diferentes picos, esto se debe a que el presente año es algo incierto por causa de la pandemia.

Figura 23. Gráfico de dispersión del año 2018-2021 sin valores outliers



Fuente: Elaboración propia (2021)

De la figura 23 podemos observar que cada cierta semana hay picos en las ventas, esto se debe a las ventas por campaña navideña y estas se repiten por los años en estudio. Además, en este gráfico se eliminó los valores outliers que se presentó en el año 2020 por la pandemia.

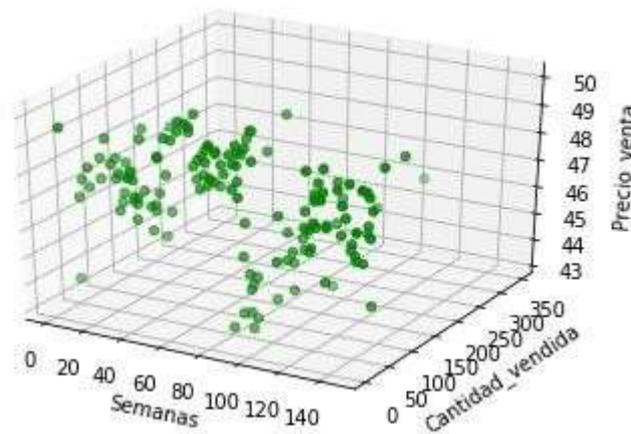
Realizamos un gráfico en 3D para ver la dispersión de las variables independientes (Semanas, P\_venta) y las dependientes (Cantidad\_v).

```
from mpl_toolkits.mplot3d import axes3d
import matplotlib.pyplot as plt

# Creamos la figura
fig = plt.figure()

# Agrrgamos un plano 3D
ax = fig.add_subplot(111,projection='3d')
x=dataset.iloc[:, 7].values
y=dataset.iloc[:, 9].values
z=dataset.iloc[:, 8].values
ax.scatter(x, y, z, c='g', marker='o')
plt.xlabel('Semanas')
plt.ylabel('Cantidad_vendida')
ax.set_zlabel('Precio_venta')
plt.show()
```

Figura 24. Gráfico de dispersión en 3D

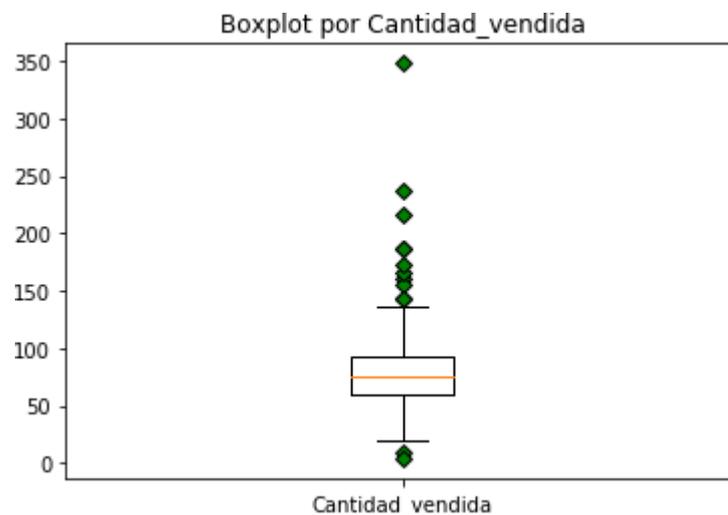


Fuente: Elaboración propia (2021)

Realizamos un gráfico tipo botplox para poder determinar si nuestra data presenta valores outliers.

```
green_diamond = dict(markerfacecolor='g', marker='D')
fig, ax = plt.subplots()
ax.set_title('Boxplot por Cantidad_vendida')
x=dataset.iloc[:, 9].values
ax.boxplot(x, flierprops=green_diamond, labels=["Cantidad_vendida"])
```

Figura 25. Gráfico tipo botplox con valores outliers



Fuente: Elaboración propia (2021)

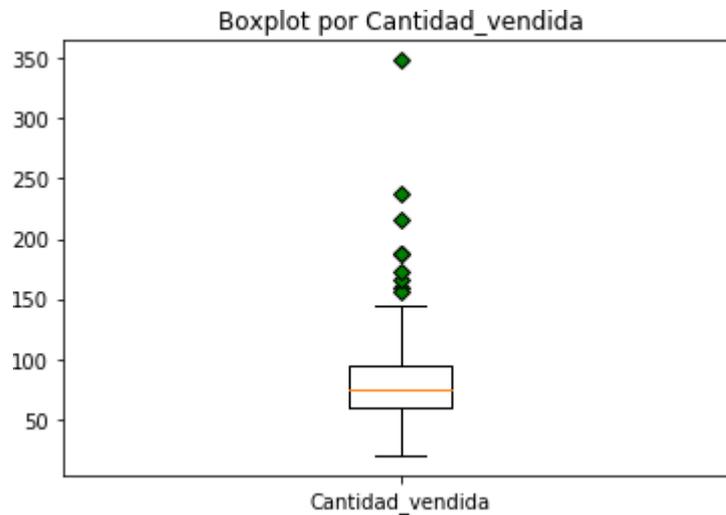
Realizamos un gráfico tipo botplox eliminando los outliers de nivel inferior, estos valores se dieron ya que eran semanas de inicio de pandemia.

```

green_diamond = dict(markerfacecolor='g', marker='D')
fig, ax = plt.subplots()
ax.set_title('Boxplot por Cantidad_vendida')
x=datasetso.iloc[:, 9].values
ax.boxplot(x, flierprops=green_diamond, labels=["Cantidad_vendida"])

```

Figura 26. Gráfico tipo botplox sin valores outliers



Fuente: Elaboración propia (2021)

En este paso separamos las variables independientes de las dependientes por cada año.

```

X = dataset[['Semana_cont', 'P_venta']]
y = dataset['Cantidad_v']
X18 = dataset2018[['Semana', 'P_venta']]
y18 = dataset2018['Cantidad_v']
X19 = dataset2019[['Semana', 'P_venta']]
y19 = dataset2019['Cantidad_v']
X20 = dataset2020[['Semana', 'P_venta']]
y20 = dataset2020['Cantidad_v']
X21 = dataset2021[['Semana', 'P_venta']]
y21 = dataset2021['Cantidad_v']
Xso = datasetso[['Semana_cont', 'P_venta']]
yso = datasetso['Cantidad_v']

```

De sklearn importamos el modelo `train_test_split`, que nos ayudará a dividir nuestra data en dos conjuntos, un conjunto de entrenamiento y el otro de prueba. Se realiza la operación cada año.

```

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
X18_train, X18_test, y18_train, y18_test = train_test_split(X18, y18, test_size=0.2, random_state=0)
X19_train, X19_test, y19_train, y19_test = train_test_split(X19, y19, test_size=0.2, random_state=0)
X20_train, X20_test, y20_train, y20_test = train_test_split(X20, y20, test_size=0.2, random_state=0)
X21_train, X21_test, y21_train, y21_test = train_test_split(X21, y21, test_size=0.2, random_state=0)
Xso_train, Xso_test, yso_train, yso_test = train_test_split(Xso, yso, test_size=0.2, random_state=0)

```

De sklearn importamos el modelo de regresión lineal, modelo usado para el presente trabajo, y mandamos a entrenar el modelo por cada año.

```
from sklearn.linear_model import LinearRegression
modelo = LinearRegression()
modelo.fit(X_train, y_train)
modelo18 = LinearRegression()
modelo18.fit(X18_train, y18_train)
modelo19 = LinearRegression()
modelo19.fit(X19_train, y19_train)
modelo20 = LinearRegression()
modelo20.fit(X20_train, y20_train)
modelo21 = LinearRegression()
modelo21.fit(X21_train, y21_train)
modeloso = LinearRegression()
modeloso.fit(Xso_train, yso_train)
```

Calculamos el coeficiente de regresión por cada modelo

```
print('2018-2021:',pd.DataFrame(modelo.coef_, X.columns, columns=['Coefficient']))
print('2018:',pd.DataFrame(modelo18.coef_, X.columns, columns=['Coefficient']))
print('2019:',pd.DataFrame(modelo19.coef_, X.columns, columns=['Coefficient']))
print('2020:',pd.DataFrame(modelo20.coef_, X.columns, columns=['Coefficient']))
print('2021:',pd.DataFrame(modelo21.coef_, X.columns, columns=['Coefficient']))
print('2018-2021 SO:',pd.DataFrame(modeloso.coef_, X.columns, columns=['Coefficient']))
```

```
2018-2021:          Coefficient
Semana_cont      -0.018614
P_venta           1.799170
2018:             Coefficient
Semana_cont       8.520870
P_venta           3.914961
2019:             Coefficient
Semana_cont       1.150881
P_venta           0.097180
2020:             Coefficient
Semana_cont       2.175751
P_venta          12.329698
2021:             Coefficient
Semana_cont      -0.378157
P_venta           3.395169
2018-2021 SO:          Coefficient
Semana_cont      -0.024983
P_venta          -0.554992
```

Realizamos la comparación entre las ventas reales y las ventas predichas para el modelo desarrollado (Ver tabla 21, tabla 22, tabla 23 , tabla 24, tabla 25, tabla 26).

```

y_pred = modelo.predict(X_test)
y18_pred = modelo18.predict(X18_test)
y19_pred = modelo19.predict(X19_test)
y20_pred = modelo20.predict(X20_test)
y21_pred = modelo21.predict(X21_test)

```

```

df = pd.DataFrame({'Ventas Actuales': y_test, 'Ventas Predecidas': y_pred})
df

```

Tabla 21. Cuadro comparativo de las ventas actuales vs ventas predichas 2018-2021

	Ventas Actuales	Ventas Predecidas
7	70	82.374052
24	60	82.850617
27	66	81.799224
120	34	82.983475
143	69	80.062199
51	111	82.515003
43	83	84.787341
22	50	81.612003
112	62	84.829802
26	70	81.966080
40	55	84.042695
78	10	84.405558
131	187	84.451094
62	84	84.374977
71	75	84.506816
91	44	78.347261

Fuente: Elaboración propia (2021)

```

df18 = pd.DataFrame({'Ventas Actuales': y18_test, 'Ventas Predecidas': y18_pred})
df18

```

Tabla 22. Cuadro comparativo de las ventas actuales vs ventas predichas del año 2018

	Ventas Actuales	Ventas Predecidas
8	75	101.071252
6	49	80.855250
4	59	60.557976

Fuente: Elaboración propia (2021)

```
df19 = pd.DataFrame({'Ventas Actuales': y19_test, 'Ventas Predecidas': y19_pred})
df19
```

Tabla 23. Cuadro comparativo de las ventas actuales vs ventas predichas del año 2019

	Ventas Actuales	Ventas Predecidas
43	112	105.916914
32	94	93.464734
38	88	100.271655
33	83	94.637184
11	64	69.114385
2	66	58.749529
31	98	92.292401
41	87	103.676123
29	83	90.003729
26	55	86.486796
4	74	60.995566

Fuente: Elaboración propia (2021)

```
df20 = pd.DataFrame({'Ventas Actuales': y20_test, 'Ventas Predecidas': y20_pred})
df20
```

Tabla 24. Cuadro comparativo de las ventas actuales vs ventas predichas del año 2020

	Ventas Actuales	Ventas Predecidas
22	44	96.355647
20	55	83.184986
25	35	87.305436
4	75	74.635523
10	56	76.028859
15	48	113.251104
28	96	122.390199
11	10	88.319043

Fuente: Elaboración propia (2021)

```
df21 = pd.DataFrame({'Ventas Actuales': y21_test, 'Ventas Predecidas': y21_pred})
df21
```

Tabla 25. Cuadro comparativo de las ventas actuales vs ventas predichas del año 2021

	Ventas Actuales	Ventas Predecidas
22	84	68.276951
20	72	71.193298
25	52	73.787341
4	63	77.719550
10	62	75.828847
15	50	77.843691
28	67	66.228856
11	48	73.448410

Fuente: Elaboración propia (2021)

```
dfso = pd.DataFrame({'Ventas Actuales': yso_test, 'Ventas Predecidas': yso_pred})
dfso
```

Tabla 26. Cuadro comparativo de las ventas actuales vs ventas predichas 2018-2021 sin valores outliers

	Ventas Actuales	Ventas Predecidas
7	70	82.374052
24	60	82.850617
27	66	81.799224
66	130	84.457434
73	73	84.206690
98	75	80.924801
62	84	84.374977
22	50	81.612003
111	83	84.147795
26	70	81.966080
40	55	84.042695
44	112	85.187545
129	187	84.440539
60	89	83.077210
136	113	84.924058
37	94	81.378748
16	66	82.599617

Fuente: Elaboración propia (2021)

Para poder evaluar el modelo se usaron las métricas MAE, MSE y RMSE, y obtenemos los distintos resultados por cada modelo realizado.

```

from sklearn import metrics
print('Mean Absolute Error 2018-2021:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error 2018-2021:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error 2018-2021:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('Mean Absolute Error 2018:', metrics.mean_absolute_error(y18_test, y18_pred))
print('Mean Squared Error 2018:', metrics.mean_squared_error(y18_test, y18_pred))
print('Root Mean Squared Error 2018:', np.sqrt(metrics.mean_squared_error(y18_test, y18_pred)))
print('Mean Absolute Error 2019:', metrics.mean_absolute_error(y19_test, y19_pred))
print('Mean Squared Error 2019:', metrics.mean_squared_error(y19_test, y19_pred))
print('Root Mean Squared Error 2019:', np.sqrt(metrics.mean_squared_error(y19_test, y19_pred)))
print('Mean Absolute Error 2020:', metrics.mean_absolute_error(y20_test, y20_pred))
print('Mean Squared Error 2020:', metrics.mean_squared_error(y20_test, y20_pred))
print('Root Mean Squared Error 2020:', np.sqrt(metrics.mean_squared_error(y20_test, y20_pred)))
print('Mean Absolute Error 2021:', metrics.mean_absolute_error(y21_test, y21_pred))
print('Mean Squared Error 2021:', metrics.mean_squared_error(y21_test, y21_pred))
print('Root Mean Squared Error 2021:', np.sqrt(metrics.mean_squared_error(y21_test, y21_pred)))
print('Mean Absolute Error 2018-2021 SO:', metrics.mean_absolute_error(yso_test, yso_pred))
print('Mean Squared Error 2018-2021 SO:', metrics.mean_squared_error(yso_test, yso_pred))
print('Root Mean Squared Error 2018-2021 SO:', np.sqrt(metrics.mean_squared_error(yso_test, yso_pred)))

```

```

Mean Absolute Error 2018-2021: 33.80585228392219
Mean Squared Error 2018-2021: 3489.071173959865
Root Mean Squared Error 2018-2021: 59.06836017666196
Mean Absolute Error 2018: 19.828159059593037
Mean Squared Error 2018: 565.6314642056279
Root Mean Squared Error 2018: 23.78300788810423
Mean Absolute Error 2019: 10.615520840279556
Mean Squared Error 2019: 174.75362781997717
Root Mean Squared Error 2019: 13.219441282443716
Mean Absolute Error 2020: 40.39996873742851
Mean Squared Error 2020: 2220.0844243455867
Root Mean Squared Error 2020: 47.11777185251428
Mean Absolute Error 2021: 15.116091816466803
Mean Squared Error 2021: 319.2428536064839
Root Mean Squared Error 2021: 17.867368401823587
Mean Absolute Error 2018-2021 SO: 21.92129112385876
Mean Squared Error 2018-2021 SO: 898.5669373611786
Root Mean Squared Error 2018-2021 SO: 29.97610610738457

```

## **CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES**

### **6.1. Conclusiones**

En el presente trabajo se implementó la técnica de regresión lineal múltiple para predecir el volumen de ventas de pantalones de caballeros de la sede Cusco de la empresa Brusko S.AC.

Con respecto a la metodología empleada, en la primera etapa se realizó la recolección de datos de ventas por semana del periodo Oct-2018 a Oct 2021. A esta data se le aplicó una limpieza de datos de manera manual (ejm: se eliminaron las semanas que no tenían datos).

En la segunda etapa, de la data ya depurada, se realizó la predicción bajo dos escenarios: predicción anual y predicción acumulada. En el primer escenario, se llegó a obtener un modelo que para el año 2018 predice con un RMSE de 24.78 error; para el año 2019, con 13.22; para el 2020 con 47.12 y para el 2021 con 17.87. Para el segundo escenario, el modelo predice con un error de 29.98.

En la última etapa, se validaron, evaluaron e interpretaron los modelos obtenidos para ambos escenarios. El modelo que se ha tomado es el que ha producido menor error; es decir el del año 2019 con un RMSE de 13.22.

Es importante recalcar que el error hallado considera dos variables (fechas y precio de venta). Sin embargo, este depende significativamente de otros factores externos los cuales no han sido considerados para el modelo tal y como se señala en la delimitación de esta investigación.

### **6.2. Recomendaciones**

Se sugiere para las futuras investigaciones que impliquen predicción de ventas, obtener la mayor cantidad de data posible, así como emplear varias variables (ejm: grupo etario de los clientes, cantidad de habitantes en la zona, modelo del producto, stock de tienda, etc.) para con ello lograr un menor error y por ende una demanda más acertada.

Si se desea hacer un análisis de todos los productos que vende una empresa se debe tener data con respecto a las características de dichos productos como marca, modelo, material,

etc., y estas deben ser validadas por un especialista para ver la calidad de los datos que se usen para el modelo del pronóstico.

Finalmente, se pueden emplear distintas técnicas de predicción además de la regresión lineal, en específico aquellas que consideren temporalidad de la data como Arima y sus variantes, para conocer el comportamiento de la variable en estudio a través del tiempo, para lo cual se tienen que estudiar periodos y datos que sean lo más homogéneos, y así poder realizar las predicciones. No se utilizó dicha técnica para nuestro trabajo de investigación porque dicho tópico no estuvo considerado en el syllabus del curso.

## REFERENCIAS BIBLIOGRÁFICAS

Barrientos, E., y Mamani, S. (2019). *Modelos de aprendizaje supervisado como apoyo a la toma de decisiones en las organizaciones basados en datos de redes sociales: Una revisión sistemática de la literatura*. [Tesis para optar el grado académico de Ingeniería de Sistemas, Universidad Peruana Unión]. Repositorio de la Universidad Peruana Unión. [https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/2544/Estrella\\_Trabajo\\_Bachillerato\\_2019.pdf?sequence=1&isAllowed=y](https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/2544/Estrella_Trabajo_Bachillerato_2019.pdf?sequence=1&isAllowed=y)

Beunza, J., Puertas, E., y Condés, E. (2019). *Manual práctico de Inteligencia Artificial en entornos sanitarios*. Elsevier España. <https://books.google.com.pe/books?id=88nSDwAAQBAJ&pg=PA35&dq=aprendizaje+supervisado+machine+learning&hl=es&sa=X&ved=2ahUKEwj90bDen4D0AhWeGbkGHey7A0U4FBC7BXoECAkQBg#v=onepage&q=aprendizaje%20supervisado%20machine%20learning&f=false>

Bobadilla, J. (2020). *Machine Learning y Deep Learning Usando Python, Scikit y Keras*. Rama. [https://books.google.com.pe/books?id=iAAyEAAAQBAJ&pg=PA13&dq=Machine+Learning&hl=es&sa=X&ved=2ahUKEwjlsJmuyf\\_zAhUUE7kGHfT8DKkQ6AF6BAgEEAI#v=onepage&q=Machine%20Learning&f=false](https://books.google.com.pe/books?id=iAAyEAAAQBAJ&pg=PA13&dq=Machine+Learning&hl=es&sa=X&ved=2ahUKEwjlsJmuyf_zAhUUE7kGHfT8DKkQ6AF6BAgEEAI#v=onepage&q=Machine%20Learning&f=false)

Cancino, C. (2012). Matriz de Análisis Foda cuantitativo. (documento en proceso) Universidad de Chile. <https://gestion.pe/economia/ventas-del-sector-retail-en-peru-crecieron-88-en-primer-trimestre-afirma-gfk-nndc-noticia/?ref=gesr>

Cano, J. (2021), Colombia. *Desarrollo de Solución Analítica para la predicción de la demanda de línea*. Universidad de Antioquia.

Céspedes, A. (2017), en Santiago, Chile, realizó la investigación titulada: *Construcción De Modelo De Forecast Para Estimación De Demanda En Una Empresa Multinacional De Retail*, Memoria De Titulación Para Optar Al Título De Ingeniero Civil Informático. Recuperado de <https://repositorio.usm.cl/bitstream/handle/11673/41250/3560902038636UTFSM.pdf?sequence=1&isAllowed=y>

*Exportaciones textiles crecerían entre 10% a 15% este año superando niveles prepandemia, según proyecta CCL* (31 de agosto de 2021). Diario El Comercio. Recuperado el 12 de octubre de 2021 de <https://elcomercio.pe/economia/exportaciones-textiles-crecerian-entre-10-a-15-este-ano-superando-niveles-prepandemia-segun-proyecta-ccl-nndc-noticia/>

*75% de las empresas peruanas indican que la Inteligencia Artificial y la automatización son los temas centrales de los directorios* (20 de julio del 2018). EY. Recuperado el 04 de noviembre de 2021 de:

[https://www.ey.com/es\\_pe/news/2018/07/75-empresas-peruanas-indican-inteligencia-artificial-automatizac](https://www.ey.com/es_pe/news/2018/07/75-empresas-peruanas-indican-inteligencia-artificial-automatizac)

Fabian, J. (2021). Clase de machine learning aplicado a los negocios, curso de actualización de la universidad ESAN 2021-2.

Gamarra, F. (2019). Modelo basado en Machine Learning para el neurorendimiento académico de estudiantes universitarios. *Ciencia y Tecnología*, 5, 10-18. <https://revistas.ujcm.edu.pe/index.php/rctd/article/view/137/121>

Garcete, A., Benítez, R., Pinto-Roa, D., y Vázquez, A. (2017). *Técnica de pronóstico de la demanda basada en Business Intelligence y Machine Learning. Simposio Argentino sobre Tecnología y Sociedad (STS)*. Pag. 193-202

González, A. y Alba, F. (s.f.). Machine learning en la Industria: El caso de la siderurgia. *Universidad de La Rioja*, 55-63. <https://www.mincotur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/405/GONZALEZ%20MARCOS%20Y%20ALBA%20EL%20C3%8DAS.pdf>

Harrington, P. (2012). *Machine Learning In Action*. Manning. [http://www2.ift.ulaval.ca/~chaib/IFT-4102-7025/public\\_html/Fichiers/Machine\\_Learning\\_in\\_Action.pdf](http://www2.ift.ulaval.ca/~chaib/IFT-4102-7025/public_html/Fichiers/Machine_Learning_in_Action.pdf)

León, E. (2021). *Te contamos qué es lenguaje R*. Baoss Analytics Everywhere. Recuperado el 20 de noviembre de <https://www.baoss.es/te-contamos-que-es-el-lenguaje-r/>

Luján, A. (2017). *Mejora de la gestión de pronósticos de la demanda para reducir los inventarios en una empresa textil*. [Tesis para optar el título profesional de Licenciado en Ingeniería Industrial y Comercial, Universidad San Ignacio de Loyola]. Repositorio de la Universidad San Ignacio de Loyola. [http://repositorio.usil.edu.pe/bitstream/USIL/3475/1/2017\\_Lujan-Arellano.pdf](http://repositorio.usil.edu.pe/bitstream/USIL/3475/1/2017_Lujan-Arellano.pdf)

Manrique, E. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas y Tecnologías de Información*, Vol<sup>o</sup>4 (N<sup>o</sup>E28), 586-599. <https://www.proquest.com/openview/c7e24c997199215aa26a39107dd2fe98/1?pq-origsite=gscholar&cbl=1006393>

Méndez, C. (1995). *Metodología, guía para elaborar diseños de investigación en ciencias económicas, contables y administrativas*. McGraw-Hill. pág. 92.

Murillo, J. (2011). Métodos de investigación de enfoque experimental. Recuperado el 1 de noviembre de <https://www.postgradoune.edu.pe/pdf/documentos-academicos/ciencias-de-la-educacion/10.pdf>

Ortega, A. (2018). Enfoque de investigación. Recuperado el 1 de noviembre de [https://www.researchgate.net/profile/Alfredo-Otero-Ortega/publication/326905435\\_ENFOQUES\\_DE\\_INVESTIGACION/links/5b6b7f9992851ca650526dfd/ENFOQUES-DE-INVESTIGACION.pdf](https://www.researchgate.net/profile/Alfredo-Otero-Ortega/publication/326905435_ENFOQUES_DE_INVESTIGACION/links/5b6b7f9992851ca650526dfd/ENFOQUES-DE-INVESTIGACION.pdf)

Pedrosa, S. (06 de septiembre del 2016). *Curva de demanda*. Recuperado el 28 de noviembre del 2021 de: <https://economipedia.com/definiciones/curva-de-demanda.html>

Peiró, A. (2015). *Demanda*. Economipedia. Recuperado el 12 de octubre de <https://economipedia.com/definiciones/demanda.html>

Pérez, R. (2019), Lima, Perú. *Generación de reglas de asociación para productos de retail utilizando el algoritmo FP-Growth paralelo*. Universidad de Lima

Russell, S. y Norvig, P. (2008). *Inteligencia Artificial Un Enfoque Moderno*. Pearson Education. Recuperado el 08 de noviembre del 2021 de: <https://luismejias21.files.wordpress.com/2017/09/inteligencia-artificial-un-enfoque-moderno-stuart-j-russell.pdf>

Rusu, C. (2011). Metodología de la investigación. *Universidad Cesar Vallejo*. Recuperado el 1 de noviembre de 2021 de: [http://www.formaciondocente.com.mx/06\\_RinconInvestigacion/01\\_Documentos/El%20Alcance%20de%20la%20Investigacion.pdf](http://www.formaciondocente.com.mx/06_RinconInvestigacion/01_Documentos/El%20Alcance%20de%20la%20Investigacion.pdf)

Sittón, I., Rodríguez, S. y Muñoz, L. (2018). Diseño de un modelo predictivo en el contexto industria 4.0. *Knowledge*, 2018, 543-551. <https://knepublishing.com/index.php/KnE-Engineering/article/view/1458>

Yu-Sun, L., Chun-Hung, H., & Shingchern, Y. (2021). Estimation of channel MSE for ATSC 3.0 receiver and its applications. *ScienceDirect*, 1-5. <https://reader.elsevier.com/reader/sd/pii/S2405959521001132?token=093D03572BEF8BB95BE57FBCB130D4AB476EE0CFE607B51ED81F2292D20269232EE7CBD7178E9423B704A0AC0960303A&originRegion=us-east-1&originCreation=20211201012013>

---

## ANEXOS

### Entrevista a la analista de ventas de la Corporación Brusko S.A.C.

#### Detalles de la entrevista

Nombre

de la empresa: Brusko S.A.C Fecha: 12/11/2021 Hora: 4:00 pm

Nombre del

entrevistador: Ronald Ccoyccosi

Nombre del

entrevistado: Brisa Salazar

#### Preguntas que debe realizar el entrevistador

Pregunta N°1: ¿Qué tipo de productos comercializan?

Respuesta: La empresa es un retail de prendas de vestir para toda la familia. Tenemos prendas para todos los usuarios: caballero, dama, jóvenes y niños.

Pregunta N°2: ¿Cuál es su producto o productos que más se venden?

Respuesta: Los productos que más se venden son polos y pantalones de todos los usuarios.

Pregunta N°3: ¿En qué provincia tienen mayor volumen de ventas?

Respuesta: La mayor comercialización de productos se realiza en la tienda de Cusco.

Pregunta N°4: ¿Cómo hacen para saber qué cantidad deben de tener en stock para la reposición en las tiendas?

Respuesta: Los supervisores de cada tienda tienen la responsabilidad de revisar el stock disponible en cada tienda y notificar al almacén en caso se necesite reposición de cierto producto. Estas notificaciones normalmente se realizan 1 vez por semana.

Pregunta N°5: ¿Quién brinda la información de que tipos y cuántos productos necesitan reponer en la tienda?

Respuesta: La cantidad a enviada a cada tienda se estima con la cantidad de ventas de cada punto de venta; si una tienda está teniendo un mejor desempeño que las otras empieza a recibir mayor cantidad de productos.

---

Pregunta N°6: ¿Cuánto tiempo en promedio demoran en llegar sus productos de su almacén a las tiendas?

---

Respuesta: Los tiempos de entrega varían dependiendo a la distancia de las tiendas, este tiempo puede ir de un par de días a 2 semanas.

---

Pregunta N°7: ¿Cuál es la frecuencia de devolución de productos que tienen en sus tiendas?

---

Respuesta: La cantidad de devoluciones es por parte de los establecimientos es mínima por lo que no se toma un registro riguroso de estas cantidades.

---

Pregunta N°8: ¿Qué productos se venden más con respecto a la categoría género?

---

Respuesta: Con respecto a los géneros; los productos más vendidos son los polos por mayor cantidad de modelos.

---

Pregunta N°9: En los últimos años ¿Qué factores externos a la empresa han identificado que han perjudicado las ventas de sus productos?

---

Respuesta: En los últimos años, los eventos que más han afectado a la empresa fueron las restricciones por cuarentena, subida del dólar y protestas con bloqueo de carreteras.

---

### **Notas adicionales**

Los productos de los cuales se cuenta con más detalle en este momento son los pantalones de caballero debido a que es necesario un mejor control de los productos que generan más margen.