TECNICATURA SUPERIOR EN CIENCIAS DE DATOS E INTELIGENCIA ARTIFICIAL

Estadística y exploración de datos II
Ciencia de Datos-II

Aprendizaje Basado en Proyectos

Análisis exhaustivo del dataset

“pet_store_records_2020.csv”

Integrantes del Grupo:

- Cáceres Giménez, Cesia Fiorella
- Di Campli, Gastón
- Lorenzati, Valentino
- Menón, Nicolas
- Terreno, Alejo

Profesores a Cargo: Nahuel Pratta/Marcos Augusto Ugarte

Tipo de Proyecto: Tecnológico y de Investigación

Octubre – 2025

Dirección General de EDUCACIÓN TÉCNICA Y FORMACIÓN PROFESIONAL
Ministerio de EDUCACIÓN
CÓRDOBA Seguimos haciendo

Índice

EJES | UNIDADES CONCEPTUALES:

Los ejes temáticos de este proyecto contemplan todo lo visto en las materias Ciencia de Datos II y Estadística y Exploración de Datos II del cuarto cuatrimestre de la carrera Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial:

- Ciencia de datos II: Numpy, Pandas, Introducción al Machine Learning, Correlación, Regresión lineal, lineal múltiple y logística.
- Estadística II: Estimación de parámetros, contraste de hipótesis, Test de ANOVA, correlación, Regresión lineal, lineal múltiple y logística.

PROBLEMÁTICA | NECESIDAD | CASO:

El caso de estudio desarrollado a lo largo de las evidencias de aprendizaje y este proyecto consta en, a partir de un dataset, aplicar las técnicas estadísticas y los algoritmos estudiados para la obtencion de informacion a traves de los datos, así como la el uso de los mismos para la inferencia de escenarios.

FUNDAMENTACIÓN | HIPÓTESIS

El presente proyecto surge de la necesidad de aplicar de manera integrada los conocimientos de Estadística II y Ciencia de Datos II en un contexto real, utilizando un dataset de ventas de una tienda de mascotas. Este conjunto de datos representa un entorno comercial concreto en el que convergen diversas variables (como productos, precios, categorías, fechas, cantidades vendidas), lo que permite analizar fenómenos económicos, de consumo y de comportamiento del cliente desde una perspectiva cuantitativa y predictiva.

En la actualidad, las organizaciones necesitan tomar decisiones fundamentadas en datos. El análisis estadístico y el aprendizaje automático ofrecen herramientas para identificar patrones de compra, medir la efectividad de estrategias de venta y predecir comportamientos futuros del mercado. En el caso

del petshop, comprender cómo influyen factores como el precio o la categoría del producto en las ventas permite optimizar recursos, mejorar la rentabilidad y lograr un conocimiento profundo del negocio en general para ayudar a tomar decisiones estratégicas.

La importancia del proyecto radica en su aplicabilidad práctica, los resultados no solo son útiles para el contexto educativo, sino también para cualquier negocio minorista que gestione inventarios y ventas. A través del uso de las herramientas antes mencionadas se logra automatizar el tratamiento de datos, explorar correlaciones, generar modelos de regresión lineal y logística, y evaluar hipótesis sobre los factores que más inciden en las ventas.

OBJETIVO GENERAL

Analizar y modelar el comportamiento de las ventas de un petshop mediante la aplicación de herramientas estadísticas y técnicas de ciencia de datos, con el propósito de identificar los factores que influyen en las ventas, validar hipótesis sobre las relaciones entre variables y desarrollar modelos predictivos que aporten valor al proceso de toma de decisiones comerciales.

OBJETIVOS ESPECÍFICOS

- **Explorar y depurar el dataset** de ventas de la tienda de mascotas, identificando valores faltantes, atípicos o inconsistencias, para asegurar la calidad de los datos utilizados en el análisis.
- **Realizar un análisis descriptivo** de las variables numéricas y categóricas, empleando medidas de tendencia central, dispersión y gráficos exploratorios que permitan comprender la estructura general de los datos.
- **Evaluar la normalidad y homocedasticidad** de las variables de interés mediante pruebas estadísticas (como Shapiro-Wilk, Levene y QQ plots) para determinar la aplicabilidad de modelos paramétricos.
- **Aplicar análisis de varianza (ANOVA)** de una y dos vías para contrastar hipótesis sobre las diferencias en los montos de venta según categorías de producto y tipo de mascota.

- **Analizar la correlación entre variables cuantitativas** (por ejemplo, precio y monto de venta) utilizando coeficientes de correlación adecuados para identificar relaciones significativas.
- **Desarrollar y ajustar modelos de regresión lineal y múltiple** para modelar el comportamiento de las ventas y predecir resultados en función de factores relevantes.
- **Interpretar los resultados obtenidos** desde una perspectiva comercial, proponiendo conclusiones y recomendaciones basadas en evidencia estadística y en los modelos predictivos desarrollados.
- **Integrar los conocimientos teóricos de estadística y ciencia de datos** en una aplicación práctica que refleje la utilidad de estas herramientas para la toma de decisiones estratégicas en un entorno de negocios.

CRONOGRAMA

Delimitado por la duración del proyecto estudiantil y relacionado con las actividades propias del mismo.

Objetivo Específico	Acciones	Recursos	Tiempos Estimados
Explorar y depurar el dataset	- Importar el dataset desde CSV. - Identificar valores nulos, duplicados o inconsistentes. - Normalizar formatos de datos y eliminar registros atípicos no representativos.	Python (Pandas, Numpy), Google Colab, Dataset Kaggle.	Semana 1

Realizar un análisis descriptivo	- Calcular estadísticas descriptivas (media, mediana, desviación estándar). - Generar visualizaciones (boxplots, histogramas, gráficos de barras). - Describir patrones generales y distribución de variables.	Pandas, Matplotlib, Seaborn.	Semana 2
Evaluar la normalidad y homocedasticidad	- Aplicar test de Shapiro-Wilk y Levene. - Utilizar QQ-plots para contrastar visualmente la normalidad. - Determinar la pertinencia de pruebas paramétricas.	Scipy, Statsmodels, Matplotlib.	Semana 2
Aplicar análisis de varianza (ANOVA)	- Realizar ANOVA de una vía para comparar montos de venta por categoría. - Aplicar ANOVA de dos vías considerando tipo de mascota y categoría. - Interpretar los valores p y diferencias significativas.	Statsmodels, Scipy.	Semana 3
Analizar la correlación entre variables cuantitativas	- Calcular coeficientes de correlación (Pearson o Spearman según distribución).	Pandas, Seaborn (heatmap).	Semana 3

	- Graficar matrices de correlación para visualizar relaciones entre variables.		
Desarrollar y ajustar modelos de regresión lineal y múltiple	- Seleccionar variables predictoras relevantes. - Ajustar modelos lineales y evaluar R^2 y significancia de los coeficientes. - Validar supuestos de los modelos (residuos, colinealidad).	Statsmodels, Scikit-learn, Python.	Semana 4
Interpretar los resultados obtenidos	- Analizar los hallazgos estadísticos en relación con el contexto comercial. - Redactar conclusiones sobre los factores que más influyen en las ventas. - Traducir los resultados técnicos a un lenguaje gerencial.	Informe final, gráficos de apoyo, conocimientos teóricos.	Semana 5
Integrar los conocimientos teóricos de estadística y ciencia de datos	- Sistematizar los resultados en un documento final. - Relacionar los métodos utilizados con los contenidos de ambas materias. - Presentar la síntesis del proceso en una exposición grupal.	Google Colab, PowerPoint/Canva, documentación del equipo.	Semana 6

PRODUCTO FINAL

Repositorio: https://github.com/Grupo-11-CDIA/Ciencia-de-datos

CONCLUSIONES | RESULTADOS ESPERADOS

El desarrollo del proyecto permitió recorrer de manera integral los principales temas abordados a lo largo del cuatrimestre: desde la exploración inicial de los datos, el análisis de correlación y la aplicación del ANOVA, hasta la implementación de modelos de regresión y el diagnóstico de sus residuos. Cada instancia del trabajo fue una oportunidad para conectar los conceptos teóricos con la práctica, comprendiendo cómo las distintas herramientas estadísticas y de programación se complementan para transformar los datos en información útil y confiable.

A través de este proceso, no solo se afianzaron conocimientos técnicos como la estimación de parámetros, los contrastes de hipótesis o la interpretación de coeficientes, sino también competencias más amplias vinculadas al pensamiento crítico y la toma de decisiones basadas en evidencia. Enfrentar resultados inesperados y analizarlos con criterio fue parte del aprendizaje, reforzando la importancia de interpretar más allá de los números.

En definitiva, este proyecto integró el trabajo colaborativo, la interpretación analítica y la aplicación práctica de técnicas de modelado, consolidando las capacidades necesarias para el ejercicio profesional en el ámbito del análisis de datos y la ciencia aplicada.

RESEÑA GRUPAL SEGUNDO CUATRIMESTRE

Durante el semestre anterior no trabajamos con el mismo grupo de trabajo: Cesia y Nicolas trabajaron juntos en un proyecto sobre la Oferta Virtual de carreras universitarias de grado; y Gaston, Valentino y Alejo en un proyecto sobre estadísticas de accidentes viales. Si bien no trabajamos con el mismo dataset ambos aplicamos técnicas y algoritmos similares.

En esta primera etapa se realizó la depuración de los datasets, eliminando valores faltantes o inconsistentes y normalizando las variables para facilitar su análisis posterior. Luego se aplicaron herramientas de Numpy y Pandas para explorar los datos y calcular medidas descriptivas como medias, medianas, modas y desviaciones estándar. Finalmente, mediante el uso de Matplotlib, se elaboraron gráficos de barras, histogramas y diagramas de dispersión, que

permitieron visualizar patrones relevantes y relaciones entre variables, sentando así las bases del análisis estadístico posterior.

En el segundo cuatrimestre, con la conformación actual del grupo y la elección del dataset aprendimos técnicas de modelado y análisis inferencial. lo que nos permitió lograr un análisis más complejo e integral de los datos y el comportamiento de los mismos.

Profundizamos en el uso de estimación de parámetros e intervalos de confianza, además de contrastes de hipótesis para verificar diferencias significativas entre categorías de productos y periodos de venta.

Asimismo, se implementaron modelos de regresión lineal, múltiple y logística, con el objetivo de predecir el comportamiento de las ventas a partir de variables como precio, categoría o cantidad vendida. Estos modelos fueron evaluados mediante métricas de desempeño y representaciones gráficas que facilitaron la interpretación de los resultados.

En conjunto, el trabajo del segundo cuatrimestre permitió integrar las herramientas estadísticas con las de machine learning, consolidando una visión más completa del proceso analítico y reforzando competencias clave del perfil profesional del Técnico Superior en Ciencia de Datos, como la interpretación crítica de la información, la toma de decisiones basadas en evidencia y la comunicación efectiva de los hallazgos.