**Universidad ORT Uruguay**

Facultad de Ingeniería

**Machine learning en producción**

**Obligatorio**

**Autores:**

Gonzalo de León - 158545

Martin Ljubicic - 154630

Joaquín Oldan - 159579

**Docentes:**

Federico Zaiter

Matías Sorozabal

**2023**

Tabla de contenido

[I. Introducción 4](#_Toc136775622)

[A. Contexto del estudio 4](#_Toc136775623)

[B. Objetivos 4](#_Toc136775624)

[C. Metodología de investigación 5](#_Toc136775625)

[II. Fundamentos de Machine Learning 6](#_Toc136775626)

[A. Definición de Machine Learning 6](#_Toc136775627)

[B. Tipos de algoritmos de Machine Learning 6](#_Toc136775628)

[C. Importancia del Machine Learning en la producción 6](#_Toc136775629)

[III. Preparación de los datos 7](#_Toc136775630)

[A. Recopilación de datos 7](#_Toc136775631)

[B. Limpieza y preprocesamiento de datos 7](#_Toc136775632)

[C. Selección de características 7](#_Toc136775633)

[IV. Modelado y entrenamiento 8](#_Toc136775634)

[A. Selección de algoritmos de Machine Learning 8](#_Toc136775635)

[B. División de conjuntos de datos 8](#_Toc136775636)

[C. Entrenamiento del modelo 8](#_Toc136775637)

[V. Implementación en producción 9](#_Toc136775638)

[A. Consideraciones de infraestructura 9](#_Toc136775639)

[B. Integración del modelo en un sistema de producción 9](#_Toc136775640)

[C. Monitoreo y mantenimiento del modelo en producción 9](#_Toc136775641)

[VI. Casos de estudio 10](#_Toc136775642)

[A. Descripción del caso de estudio 1 10](#_Toc136775643)

[B. Resultados y lecciones aprendidas del caso de estudio 1 10](#_Toc136775644)

[C. Descripción del caso de estudio 2 10](#_Toc136775645)

[D. Resultados y lecciones aprendidas del caso de estudio 2 10](#_Toc136775646)

[VII. Desafíos y recomendaciones 11](#_Toc136775647)

[A. Desafíos comunes en la implementación de Machine Learning en producción 11](#_Toc136775648)

[B. Recomendaciones para superar los desafíos 11](#_Toc136775649)

[VIII. Conclusiones 12](#_Toc136775650)

[A. Resumen de los hallazgos 12](#_Toc136775651)

[B. Contribuciones del estudio 12](#_Toc136775652)

[C. Áreas para futuras investigaciones 12](#_Toc136775653)

# I. Introducción

## A. Contexto del estudio

El Machine Learning surge como una necesidad de estudiar los datos obtenidos de cualquier ámbito para obtener información relevante para problemas planteados o para mejorar sistemas. La toma de decisiones automatizada, la automatización de procesos aporta mucho valor agregado en diferentes ámbitos.

Como se pudo observar en las presentaciones realizadas por el equipo de diversos papers, existen cada vez mas casos de uso exitosos en la materia, en la presentación de los integrantes del grupo sobre Michelangelo, se expuso que no solo es importante los resultados directos del aprendizaje, sino la metodología aplicada para llegar a ellos. Esta herramienta proporciona buenos resultados y ayudo a organizar los grupos de trabajo.

La automatización es vista como como una oportunidad de mejora para las organizaciones y permitirá redirigir sus esfuerzo a trabajos de mejora constante de estas metodologías, logrando mejorar sus procesos.

En este trabajo se buscará incursionar en estas metodologías para la implementación de un producto que pueda ser llevado a producción haciendo hincapié más en el E2E que en sus diferentes partes.

## B. Objetivos

El objetivo general del proyecto consiste en lograr un E2E de un sistema que obtenga datos de una pagina web de venta de propiedades y que sea capaz de aprender a predecir un parámetro de estas relevante, en este caso se apunta a poder distinguir la categoría del inmueble según su valor.

En principio esta variable esta directamente relacionada con el precio, aunque se podría llegar a paralelizar una categoría de valor en función de las necesidades de quien busca adquirir la vivienda. Esto ya es un desafío mas ambicioso y no será el foco del presente trabajo.

Se logran identificar objetivos mas específicos, apuntando a etapas de proceso general.

Obtener un modelo que logre analizar imágenes como datos tabulares de los elementos de estudio, si bien no hay un enfoque sobre obtener los mejores resultados, si se buscara un mínimo desempeño en esta etapa del modelo, que sirva para obtener resultados bastante aproximados del problema.

Análisis de datos, comprensión de cuales son mas relevantes y cuales no, cuales aportan mas al target que se busca aprender para luego inferir con buena precisión en datos de testeo.

Desempeño del sistema, en velocidad y escalabilidad, un buen sistema debe ser capaz de proporcionar una buena respuesta ante múltiples consultas.

Evaluar el impacto de la solución completa en cuanto a su utilidad real y la aplicabilidad a un caso real de estudio, si bien se puede considerar un objetivo secundario, seria bueno tener una referencia a su utilidad, por su valor agregado a algún proceso determinado para un caso de estudio puntual o general.

Comprender que mejoras se pueden proporcionar a futuro en la aplicación para apuntar a un mercado de clientes reales y su posibilidad de venta.

## C. Metodología de investigación

El enfoque del trabajo será con fines académicos, no se analiza un caso de estudio real, aunque luego pueda extrapolarse el trabajo, cambiando la información tratada para atender a un cliente. Se destaca igualmente que si bien no esta enfocado a un cliente, puede ser útil lo realizado en el caso de uso para cierto marco de público objetivo.

El estudio apunta a analizar información de propiedades de la página [www.gallito.com.uy](http://www.gallito.com.uy), para aprender a clasificarlas en clases de interés dependiendo su valor. Se recopilará información, se analizará y se mostraran resultados de forma amigable al usuario.

Los datos obtenidos se analizan para utilizarlos de la mejor manera en su inyección en el modelo, comprender que tan significativo resultara, y realmente si aportara valor. Datos faltantes también serán analizados para comprender la necesidad de completarlos o no, y así no influir en el aprendizaje del modelo y también luego en la inferencia en producción.

La información recopilada en el aplicativo se considera de interés general y acceso público, y no se expondrán datos personales de las cuentas que realizan las publicaciones o datos que puedan ser relacionados a ellas, será necesario revisar el dataset de imágenes y cuidar esta información para completar un trabajo dentro de los marcos éticos adecuados.

El estudio buscara respetar reglas de análisis de datos establecidos, cuidar el data leakage y el training-serving skew que puede repercutir de gran manera en los resultados y desempeños del aplicativo.

Si bien la información de la fuente es obtenida en un momento determinado del mercado de inmuebles, en definitiva, muy relacionada al momento actual del mercado, no se tienen datos históricos, se deberá cuidar las conclusiones a partir de estos datos sesgados en cierta manera.

# II. Fundamentos de Machine Learning

## A. Definición de Machine Learning

El Machine Learning, o aprendizaje automático, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y técnicas que permiten a las computadoras aprender de manera automática a través de la experiencia y los datos, sin ser programadas explícitamente. En lugar de seguir instrucciones específicas, los algoritmos de Machine Learning son capaces de reconocer patrones, identificar correlaciones y tomar decisiones o realizar predicciones basadas en los datos con los que han sido entrenados.

Este enfoque de aprendizaje automatizado ha demostrado ser eficaz en una amplia gama de aplicaciones, desde el reconocimiento de voz y la visión por computadora hasta la detección de fraudes y la personalización de recomendaciones. El objetivo del Machine Learning es permitir que las máquinas adquieran conocimiento y mejoren su rendimiento a medida que se enfrentan a nuevos datos, lo que lo convierte en una herramienta poderosa para abordar problemas complejos y tomar decisiones basadas en información.

## B. Tipos de algoritmos de Machine Learning

Existen diversos tipos de algoritmos de Machine Learning que se utilizan para abordar diferentes tipos de problemas y tareas. A continuación, algunos de los tipos más comunes de algoritmos de Machine Learning:

* Aprendizaje supervisado (Supervised Learning): En este tipo de algoritmo, se entrena al modelo utilizando un conjunto de datos de entrenamiento que incluye ejemplos etiquetados. El objetivo es que el modelo aprenda a predecir o clasificar nuevas instancias en función de las etiquetas conocidas. Los árboles de decisión están en esta categoría, muy explicativos y prácticos.
* Aprendizaje no supervisado (Unsupervised Learning): En los algoritmos de aprendizaje no supervisado, no se proporcionan etiquetas en los datos de entrenamiento. El objetivo es descubrir patrones o estructuras ocultas en los datos. Un ejemplo de estos son el clustering.
* Aprendizaje por refuerzo (Reinforcement Learning): Este tipo de algoritmo se basa en un proceso de toma de decisiones en el que un agente interactúa con un entorno y recibe recompensas o castigos en función de sus acciones. El objetivo es que el agente aprenda a tomar decisiones que maximicen las recompensas a largo plazo. Los ejemplos más característicos se atribuyen a juegos.
* Aprendizaje profundo (Deep Learning): El aprendizaje profundo utiliza redes neuronales artificiales con múltiples capas para aprender características y representaciones complejas de los datos. Estas redes pueden realizar tareas de reconocimiento de imágenes, procesamiento de lenguaje natural y traducción. Las redes CNN para tratamiento de imágenes son un ejemplo.

Este trabajo se enfocará en Deep Learning, realizando modelos de CNN + MLP para el tratamiento de imágenes y datos tabulares provistos de los datos adquiridos.

## C. Importancia del Machine Learning en la producción

Llevar un algoritmo de Machine Learning a producción es de suma importancia debido a que es el paso final para hacer uso efectivo de los modelos desarrollados y obtener beneficios reales en entornos prácticos. Como puntos importantes hay que destacar:

* Utilización efectiva de los modelos: Se puede aprovechar plenamente el valor y el potencial de los modelos desarrollados. Los modelos pueden aplicarse en tiempo real y en escala, lo que permite automatizar tareas, tomar decisiones y mejorar la eficiencia operativa.
* Toma de decisiones automatizada: Se puede tomar decisiones automatizadas basadas en datos y análisis. Esto puede conducir a una toma de decisiones más rápida, precisa y consistente, evitando errores humanos y mejorando la eficacia general de los procesos.
* Optimización de recursos: Se pueden aprovechar eficientemente los recursos disponibles. Los modelos pueden ayudar a optimizar procesos, mejorar la asignación de recursos, reducir el desperdicio y aumentar la productividad.
* Mejora continua: Permite recopilar datos en tiempo real y retroalimentar los resultados a los modelos de Machine Learning. Esto facilita la mejora continua y la capacidad de adaptación a medida que los modelos se enfrentan a nuevos escenarios y cambios en los datos de entrada.

Hay aspectos que hay que cuidar, para lograr un buen sistema!

* Estabilidad y rendimiento: Es esencial garantizar que el algoritmo funcione de manera estable y tenga un rendimiento óptimo en el entorno de producción. Esto implica considerar aspectos como el tiempo de respuesta, la escalabilidad, la eficiencia de los recursos y la capacidad de manejar grandes volúmenes de datos en tiempo real.
* Integración con la infraestructura existente: La integración adecuada del algoritmo en la infraestructura existente es crucial. Esto incluye aspectos como la compatibilidad con los sistemas y tecnologías utilizados, la gestión de datos de entrada y salida, y la seguridad y privacidad estos.
* Monitoreo y mantenimiento: Un algoritmo de Machine Learning en producción debe ser monitoreado de forma continua para asegurar que siga funcionando correctamente y proporcionando resultados precisos. También se debe realizar un mantenimiento regular para actualizar el modelo, reentrenarlo con nuevos datos y abordar posibles desviaciones.
* Evaluación y medición de resultados: Es importante establecer métricas de evaluación para medir el desempeño y el impacto del algoritmo en producción. Esto permitirá evaluar su eficacia, identificar posibles mejoras y demostrar el valor generado a los interesados.

En resumen, la implementación exitosa de un algoritmo de Machine Learning en producción es crucial para aprovechar plenamente los beneficios de los modelos desarrollados y lograr una toma de decisiones automatizada y eficiente.

# III. Preparación de los datos

## A. Recopilación de datos

## B. Limpieza y preprocesamiento de datos

## C. Selección de características

# IV. Modelado y entrenamiento

## A. Selección de algoritmos de Machine Learning

## B. División de conjuntos de datos

## C. Entrenamiento del modelo

# V. Implementación en producción

## A. Consideraciones de infraestructura

## B. Integración del modelo en un sistema de producción

## C. Monitoreo y mantenimiento del modelo en producción

# VI. Casos de estudio

## A. Descripción del caso de estudio 1

## B. Resultados y lecciones aprendidas del caso de estudio 1

## C. Descripción del caso de estudio 2

## D. Resultados y lecciones aprendidas del caso de estudio 2

# VII. Desafíos y recomendaciones

## A. Desafíos comunes en la implementación de Machine Learning en producción

## B. Recomendaciones para superar los desafíos

# VIII. Conclusiones

## A. Resumen de los hallazgos

## B. Contribuciones del estudio

## C. Áreas para futuras investigaciones

FIN DEL DOCUMENTO