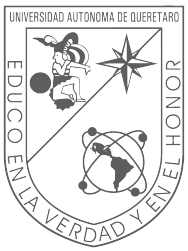


UNIVERSIDAD
**AUTÓNOMA**
DE QUERÉTARO

# Predicción de precios de la lactosa usando series de tiempo y RNN

**Proyecto final para el materia de**

**Machine Learning**

by

**Alejandra Verónica López Chiquito**

**Profesor:**
Dr Marco Antonio Aceves Fernández

Maestría en Ciencias en Inteligencia Artificial

**December 2023**

# Contents

# Abstract

Como bien se puede leer en el título de este documento, se planea implementar una red recurrente y se utilizará un conjunto de datos en forma de series de tiempo. Con esto se intentará predecir el valor del precio de la lactosa, un producto que se utiliza como materia prima en la prodcción de productos lacteos. Se tocarán temas de ciencia de datos, aprendizaje autmático y redes neuronales.

# 1 | Introduction

Dentro del ambiente del aprendizaje automático existen diferentes tipos de datos con los que se pueden trabajar, se tienen de tipo tabular, conjunto de imágenes, o donde se integran ambos tipos, es decir conjunto de datos heterogéneos.

La elección del tipo de conjunto de datos es crucial para el éxito de un proyecto de aprendizaje automático, ya que afecta directamente la elección del modelo y las técnicas de preprocesamiento de datos. Además, la ética y la privacidad deben ser consideradas al trabajar con cualquier tipo de conjunto de datos.

La predicción, en el contexto del análisis de datos y la inteligencia artificial, implica hacer estimaciones o proyecciones sobre eventos futuros basándose en datos históricos o patrones identificados. La capacidad de realizar predicciones es fundamental en una variedad de campos, desde finanzas y economía hasta ciencia, salud, meteorología y más.

La implementación de modelos de inteligencia artificial que trabajen con series de tiempo suele ser muy diferentes a los modelos de redes neuronales convencionales. Estos modelos se conoces como redes recurrentes. Estas redes tienen una arquitectura especializada para trabajar con datos secuenciales y modelar datos temporales.

En este documento se tratarán datos de tipo secuenciales, es decir, series de tiempo. ste tipo de datos es común en diversas disciplinas, y su análisis es crucial para comprender patrones temporales, identificar tendencias y realizar predicciones sobre eventos futuros.

Normalmente son registros de datos con un intervalo que no necesariamente es constante y cada dato tiene asociado una fecha asociada, frecuentamente en la que fue capturado.

# 2 | Conceptos Clave

Dentro de esta sección se explicarán dos conceptos que son fundamentales en el desarrollo de cualquier análisis de datos que es dirigido a negocios, el análisis de datos (Data Analytics)y análisis de negocio (Business Analytics). Al conocer el concepto de estos conceptos y la diferencia entre ellos se podrá aclarar los resultados de la predicción que se realizara posteriormente en este documento.

## 2.1 Data Analytics

El análisis de datos es un proceso de examinar, limpiar, transformar y modelar datos para generar información útil y ordenada que se utiliza para trazar caminos o tomar decisiones dentro de cualquier investigación o rama. Este proceso siempre involucra herramientas y técnicas para analizar gran cantidad de datos para generar predicciones, encontrar patrones u otros.

El proceso de análisis de datos se puede dividir en varias etapas, las cuales inician con identificar el problema, es decir, saber exactamente cuál es la pregunta que se quiere responder, con esto se puede construir un objetivo general. Posteriormente se requiere recopilar datos, no siempre se trata de tener la mayor cantidad de datos, más bien de tener los correctos. Una vez que se tienen los datos, el sigueinte paso sería el prepocesamiento o limpieza de datos, es decir encontrar y arreglar datos faltantes, outliers, etc., todo esto con el objetivo de buscar los mejores resultados al momento de aplicar nuestro modelo de predicción, regresión o cualquiera que sea el objetivo determinado anteriormente. Dicho de otra manera, si se ingresan datos correctos y limpios a nuestro modelo, el resultado serán más confiable. Cuando se obtienen esos resultados se requiere interpretarlos o transformarlos en información útil para el interesado, esto también se puede hacer generando gráficas u otra ayuda visual para que la comunicación de dichos resultados sea digerible al momento de presentarlos.

Existen diferentes tipos de análisis de datos, las cuales see mencionan a continuación:

- Análisis Descriptiva: Proporciona información sobre que ha pasado.
- Análisis de diagnóstico: Informa porque ha ocurrido algo
- Análisis Predictiva: Se puede saber lo que ocurrirá probablemente en el futuro.
- Análisis Presciptiva: Proporciona información sobre como actuar ante alguna situcación.

Estos tipos de análisis serán aplicados dependiendo siempre de la pregunta que se quiera responder. La pregunta se asocia siempre al paso del proceso de análisis de datos que se quiere resolver.

## 2.2 Business Analytics

Consiste en el uso de diferentes técnicas de análisis y herramientas para interpretar los datos y proporcionar información que sea relevante para la estrategia y el rendimiento empresarial. Es decir que se aplicará un análisis de datos muy similar al mencionado en la sección de data analytics con la diferencia de que la interpretación o el objetivo del business analytics is totalmente orientado a toma

de decisiones empresariales.

Las herramientas en el proceso de análisis de datos en este concepto son más sofisticadas, los resultados obtenidos tienen que tener un alto grado de confiabilidad ya que las decisiones que se toman a partir de estos resultados tienen mayor impacto.

Los tipos de Bussiness Analytics se mencionan a continuación:

- Análisis Descriptiva: Se enfoca en el seguimiento en los KPI (Key Performance Indicators) definidos por el negocio para dar seguimiento al estado actual del mismo.
- Análisis Predictiva: Tiene el objetivo de análizar las predicciones o tendencias futuras de los activos en cuestión.
- Análisis Presciptiva: Usa todos los datos pasados para generar recomendaciones para situacióne similares en el futuro.

Business Analytics se utiliza en una variedad de áreas, como finanzas, marketing, gestión de la cadena de suministro, recursos humanos y más. Al aplicar técnicas analíticas a los datos empresariales, las organizaciones pueden mejorar la eficiencia operativa, identificar oportunidades de crecimiento y mantener una ventaja competitiva en el mercado.

Los modelos que se aplican dentro de business analytics tienen pueden llegar a tener varios objetivos, como por ejemplo optimización de procesos, análisis de clientes, identificar oportunidades de crecimiento, entre otros.

## 2.3 Machine learning en el análisis de datos

Como sabemos el aprendizaje automático es una rama de la inteligencia artificial basada en sistemas pueden aprender de datos, identificar patrones con el objetivo de ayudar a tomar las mejores decisiones.

El aprendizaje automático desempeña un papel muy importante en el análisis de datos hoy en día, debido a la gran cantidad de modelos disponibles actualmente. La aplicación de técnicas de aprendizaje automático en el análisis de datos permite a las organizaciones obtener información más rápida y precisa, identificar patrones complejos y tomar decisiones más informadas basadas en datos. Este enfoque es especialmente valioso en entornos donde los conjuntos de datos son grandes y complejos.

Aquí hay algunas formas en las que el aprendizaje automático se aplica al análisis de datos

1. Predicción y Modelado:
    - Regresión: Utilizado para prever valores numéricos basados en datos históricos.
    - Clasificación: Categoriza datos en clases o etiquetas. Por ejemplo, clasificación de correos electrónicos como spam o no spam.
    - Agrupamiento (Clustering): Agrupa datos similares sin etiquetas predefinidas.
2. Detección de Anomalías: Identificación de patrones inusuales o comportamientos atípicos en datos, como la detección de fraudes en transacciones financieras.
3. Análisis de Sentimientos: Se utiliza en redes sociales, reseñas de productos, comentarios, etc., para determinar el tono emocional de un texto (positivo, negativo, neutral).
4. Procesamiento del Lenguaje Natural (PLN): Utilizado para entender y generar texto de manera natural, lo que permite análisis de texto y chatbots inteligentes.

Existen mucho otros modelos y cada uno es aplicado depeniendo del tipo de problema que se quiera resolver. Todos pueden resultar bastante útiles y pueden ser muy confiables si se les proporciona la información correcta.

# 3 | Conjunto de datos

Como un previo a este capítulo podemos mencionar que el consumo de productos en el país se considera uno de los negocios más redituables. Es por esto que el análisis de datos dentro de esta rama es impresindible, ya que las empresas dedicadas a venta de productos lacteos deben enfocarse en encontrar tendencias, patrones o características para mejorar ventas o estrategias.

Como un paso anterior a la aplicación de un modelo de inteligencia artificial, nunca debe de faltar el estudio del conjunto de datos proporcionado para resolver el problema en cuestión. Se debe tratar de obtener toda la información posible sobre los datos, no solo lo visiblemente obvio, se debe realizar un estudio más profundo, como por ejemplo con herramientas estadistica descritptiva, herramientas para generar gráficas, etc. de manera que se puedan llegar a identificar patrones, tendencias u cualquier otra característica de los datos que pueda ayudar a crear un camino en su estudio.

El conjunto de datos para este ejercicio es basado en precios de productos que se utilizan como materia prima para la producción de lácteos, como lactosa, suero, etc. Estos estan basados en una determinado periodo de tiempo, el período es semanal, es decir, que estaremos atendiento un problema de series de tiempo. En el capítulo siguiente se dará una explicación detallada sobre series de tiempo y su estudio.

## 3.1 Preprocesamiento de datos

Este capítulo está dirigido al preprocesamiento y estudio del conjunto de datos. Ya que lo que se requiere es predecir el valor de la lactosa, se tienen que conseguir el precio de la lactosa de todas las pestañas que contiene el archivo de nuestro conjunto de datos. Este archivo contiene el registro de los precios de materia prima para producir productos lacteos desde el año 1997 al año actual, 2023. El registro del precio es semanal, aunque depende el año pueden existir más registros.

Durante estre proceso se utilizaron dos librerías muy famosas para el tratamiento de datos, pandas y numpy. Se realizaron cambios dentro de los dataframes obtenidos del excel, como eliminar filas con promedios mensuales de los precios, eliminar valores nulos, etc.

Posteriormente a obtener el precio de la lactosa de todos los años en el archivo, se hará uso de la librería sklearn para escalar los datos y que el modelo pueda trabajar de manera más eficiente. El escalamiento de datos normalmente se hace entre 0 y 1. Aunque se puede escalar en diferentes rangos esto va a depender del problema que se este resolviendo. En este caso se aplicará en rango de 0 y 1 ya que los valores que estamos analizando son pequeños.

Posteriormente hablaremos de otra función diponible para "ajustar" los datos de mejor manera para que el modelo de RNN pueda trabajar mejor. Para trabajar con series de tiempo existe una función integreada en tensorflow y keras que lleva por nombre TimeseriesGenerator, esto es de gran utilidad para generar conjuntos de datos temporales. La función principal de TimeseriesGenerator es preparar datos secuenciales en lotes que se pueden utilizar para el entrenamiento de modelos.

# 4 | Series de tiempo

Las series de tiempo son conjuntos de datos que representan observaciones recopiladas o registradas secuencialmente a lo largo de un período de tiempo. Estas observaciones pueden estar espaciadas a intervalos regulares (como cada día, mes o año) o irregulares, pero la clave es que el orden temporal de las observaciones es significativo.

Como ya se mencionó anteriomente, es importante seleccionar el modelo adecuado para el tipo específico de datos y la naturaleza de la serie temporal. La validación y ajuste del modelo son cruciales para garantizar que las predicciones sean precisas y confiables. Además, la elección entre enfoques clásicos y técnicas más avanzadas dependerá del contexto y los requisitos específicos del problema.

Algunos conceptos clave relacionados con las series de tiempo incluyen:

- Tendencia: La dirección en la que los datos parecen estar moviéndose a lo largo del tiempo. Puede ser ascendente, descendente o incluso no existir.
- Estacionalidad: Patrones o ciclos que ocurren durante intervalos regulares. Esto podría ser diario, mensual, estacional, etc.
- Ciclos: Variaciones que no son de naturaleza estacional y pueden tener períodos más largos. Estos pueden ser el resultado de factores económicos, sociales u otros.
- Ruido: Variaciones aleatorias o irregulares en los datos que no siguen un patrón.
- Estacionariedad: Se considera estacionaria si sus propiedades estadísticas, como la media y la varianza, son constantes a lo largo del tiempo.

## 4.1 Modelos Predictivos

Los modelos predictivos con series de tiempo son herramientas analíticas utilizadas para predecir valores futuros basándose en patrones temporales y tendencias observadas en datos históricos. Estos modelos son muy utilizados en diversas áreas como finanzas, economía, climatología, ventas y otras disciplinas donde los datos varían con el tiempo.

A continuación se dará una breve reseña sobre algunos modelos que son utilizados para trabajar con series de tiempo, posteriormente se profundizara en uno de ellos, ya que será el cual se aplicará al conjunto de datos del problema en cuestión.

1. Media Móvil Simple (SMA): Este modelo calcula la media de un conjunto de observaciones en un período de tiempo específico. Puede ayudar a suavizar fluctuaciones a corto plazo y resaltar tendencias a largo plazo.
2. Media Móvil Exponencial (EMA): Similar a SMA, pero asigna más peso a las observaciones más recientes. Esto permite que el modelo reaccione más rápidamente a los cambios en los datos.
3. Modelos ARIMA (Autoregressive Integrated Moving Average): ARIMA es un modelo estadístico que combina componentes autoregresivos, de media móvil e integrados. Está diseñado para manejar tendencias y estacionalidades en los datos.

4. Redes Neuronales Recurrentes (RNN): Las RNN son un tipo de red neuronal diseñada para trabajar con datos secuenciales, como series temporales. Pueden aprender patrones complejos, pero también pueden requerir mucha capacidad computacional.

5. Long Short-Term Memory (LSTM): Una variante de las RNN, las LSTM son especialmente efectivas para capturar patrones a largo plazo en series temporales.

6. Prophet (Facebook): Desarrollado por Facebook, Prophet es un modelo diseñado para predecir series temporales con fuertes patrones estacionales y tendencias.

# 5 | Modelo de predicción

Las redes neuronales recurrentes (RNN) son un tipo de arquitectura de red neuronal que se utiliza comúnmente en el procesamiento de secuencias y datos temporales. A diferencia de las redes neuronales convencionales, las RNN tienen conexiones retroactivas, lo que les permite mantener y utilizar información previa durante el procesamiento de nuevas entradas.

La principal fortaleza de las RNN radica en su capacidad para trabajar con datos secuenciales o temporales, como series temporales, texto y audio. Algunas de las aplicaciones típicas de las RNN incluyen el procesamiento del lenguaje natural, la traducción automática, la generación de texto, el reconocimiento de voz y la predicción de series temporales.

## 5.1 Long short-term memory (LSTM)

Long Short-Term Memory (LSTM) es una arquitectura especializada de red neuronal recurrente (RNN), es una de las arquitecturas más utilizadas en aplicaciones que involucran datos secuenciales, como procesamiento del lenguaje natural, reconocimiento de voz y predicción de series temporales.

Las LSTMs introducen unidades de memoria especiales llamadas "celdas de memoria" y mecanismos de puertas para controlar el flujo de información en la red. Estos elementos permiten a las LSTMs retener información relevante a lo largo del tiempo y evitar problemas de desvanecimiento del gradiente que afectan a las RNN tradicionales.

Aquí hay algunas características clave de las LSTMs:

- Celdas de Memoria (Cell State): La celda de memoria es el componente central de una LSTM. Mientras que en una RNN convencional el estado oculto es la única representación interna, en una LSTM, la celda de memoria mantiene y actualiza la información a lo largo del tiempo. Actualiza su contenido combinando la información que se debe olvidar (determinada por la puerta de olvido) y la nueva información propuesta (determinada por la puerta de entrada).

- Puertas Sigmoides: Las LSTMs tienen tres puertas sigmoides que controlan la información que fluye dentro y fuera de la celda de memoria. Estas son la puerta de olvido (forget gate), la puerta de entrada (input gate), y la puerta de salida (output gate).

- Puerta de Olvido (Forget Gate): Decide qué información de la celda de memoria anterior debe ser olvidada o mantenida. La salida de esta puerta varía entre 0 (olvidar completamente) y 1 (mantener completamente).

- Puerta de Entrada (Input Gate): Determina qué nueva información debe ser almacenada en la celda de memoria. Utiliza una función sigmoide para decidir qué valores actualizados se deben agregar a la celda y una función tangente hiperbólica para generar los nuevos candidatos de celda.

- Puerta de Salida (Output Gate): Decide qué información de la celda de memoria se convertirá en la salida actual de la LSTM. Utiliza una función sigmoide para decidir cuánto de la celda de memoria se debe revelar y una función tangente hiperbólica para escalar la salida.

La estructura de puertas en las LSTMs permite el flujo selectivo de información, lo que les da la capacidad de retener y utilizar información relevante a lo largo de secuencias largas. Este diseño ha demostrado ser efectivo para abordar problemas en los que la dependencia a largo plazo es crucial.

# 6 | Resultados

Los resultados de una red neuronal recurrente (RNN) dependen en gran medida del tipo de tarea para la cual se ha entrenado la red y de la calidad de los datos utilizados. La evaluación de los resultados de una RNN implica un análisis detallado de cómo se desempeña en la tarea específica para la cual fue diseñada. Es importante considerar tanto las métricas cuantitativas como las cualitativas para obtener una comprensión completa del rendimiento del modelo.

Aunque la función TimeseriesGenerator facilita la preparación de datos temporales para el entrenamiento de modelos de aprendizaje automático, permitiendo una gestión eficiente de las series temporales en lotes, el tiempo de entrenamiento de un modelo con datos que han sido procesados con esta función se incremente considerablemente.

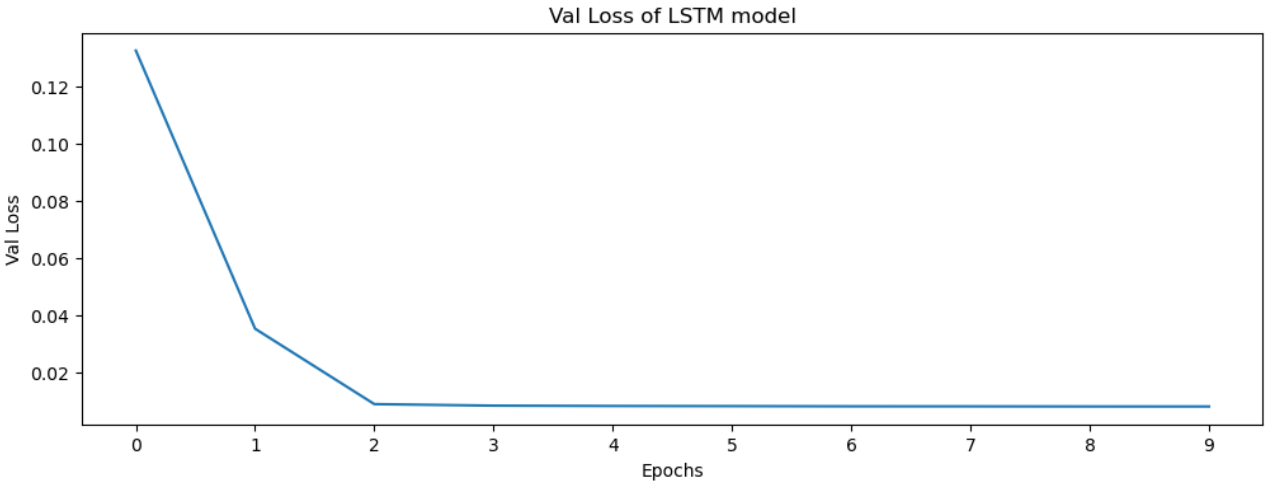


**Figure 6.1:** Comportamiento del valor de pérdida del modelo LSTM

En la figura 6.1, se puede observar el aprendizaje del modelo creado. Se entrenará el modelo con el conjunto que se definió anteriormente como batch de prueba y se evaluará el modelo con el resto del conjunto de datos.

El comportamiento del modelo al momento de las predicciones se comporto de manera correcta, aunque sin duda se podría incrementar su precisión. A continuación se muestran los resultados de diferentes métricas que se aplicarón al modelo después del entrenamiento y una vez comparadas las predicciones que hizo con lo valores reales del conjunto de datos designado para testear.

Model evaluation:

- MSE es: 0.008141545263025905
- MAE es: 0.06891899087404853
- RMSE es: 0.09023051181848579
- MAPE es: 18.105408381998238
- R2 es: 0.6181551964581371

# 7 | Conclusiones

El procesamiento de datos es una parte fundamental en la cadena de valor de la información, abarcando desde la recopilación hasta la obtención de insights valiosos.

En conclusión, las redes recurrentes (RNN) son una clase poderosa de arquitecturas de redes neuronales que han demostrado ser efectivas en el manejo de datos secuenciales y la modelización de dependencias temporales

Las Long Short-Term Memory (LSTM) han demostrado ser una herramienta poderosa en el campo del aprendizaje profundo, especialmente para el procesamiento de secuencias y la modelización de dependencias a largo plazo. Estas redes y otras arquitecturas relacionadas se han aplicado con éxito en una variedad de aplicaciones en diversas industrias. Definitivamente generar predcicciones con series de tiempo es una tarea totalmente diferente si se usaran datos comunes, ya sea tabulares o imágenes.

En general, las LSTMs han sido fundamentales para avanzar en la capacidad de las redes neuronales para manejar datos secuenciales. Sin embargo, también es importante tener en cuenta que el campo de la investigación en aprendizaje profundo está en constante evolución, y nuevas arquitecturas y enfoques continúan emergiendo para mejorar aún más el rendimiento en diversas tareas.

# Bibliography