Predicción de sólidos totales en una industria láctea mediante la aplicación de técnicas de aprendizaje automático

Delfina Berra[1][0009-0008-0931-5099]- María Della Torre[2][0009-0006-1919-963X] Mariano Ferrero[3][0000-0002-4298-6440]

[1] Universidad Nacional de Rafaela, Rafaela 2300, Santa Fe, Argentina
delfina.berra@unraf.edu.ar
[2] Universidad Nacional de Rafaela, Rafaela 2300, Santa Fe, Argentina
maria.dellatorre@unraf.edu.ar
[3] Universidad Nacional de Rafaela, Rafaela 2300, Santa Fe, Argentina
marianoferrero.mf@gmail.com

Resumen. El presente trabajo detalla la experiencia de un proyecto tecnológico llevado adelante entre una importante industria láctea en la provincia de Santa Fe y el Laboratorio de Gestión de la Información de la Universidad Nacional de Rafaela. La misma consistió en el análisis de sólidos totales en leche cruda y se llevó a cabo mediante una metodología cuantitativa tomando de base el modelo CRISP-DM. Para la etapa de comprensión de datos se realizaron reuniones entre las partes. En la instancia de análisis, se determinaron las variables a ser utilizadas y su procesamiento en modelos estadísticos. Durante el modelado, se analizaron diferentes alternativas con algoritmos de aprendizaje automático, determinando que el que mejor funcionaba era regresión lineal. Para evaluarlos se tomó de referencia el error promedio. Por último, se desarrolló una herramienta, a través de un código en el lenguaje de programación Python, adaptada a la empresa y que pudiera predecir los sólidos totales. El trabajo permitió posicionar a la Universidad como referente en tecnologías y mejora de procesos, como así también acercar a la empresa a la ciencia de datos y a tomar decisiones ágiles e informadas a partir de la reducción de tiempos operativos en la actualización de recetas.

Palabras clave: Colaboración Universidad Industria, Inteligencia Artificial, Sólidos totales en leche cruda, Modelo Predictivo.

Predicting total solids in a dairy industry by applying machine learning techniques

Abstract. This paper details the experience of a technological project carried out between a major dairy industry in the province of Santa Fe and the Information Management Laboratory of the National University of Rafaela. The project involved the analysis of total solids in raw milk and was carried out using a quantitative methodology based on the CRISP-DM model. Meetings were held between the parties for the data comprehension stage. During the analysis phase, the variables to be used were determined and processed in statistical models. During the modeling process, different alternatives were analyzed using machine learning algorithms, determining that linear regression

worked best. The mean error was used as a reference to evaluate these algorithms. Finally, a tool was developed, using Python programming code, adapted to the company and capable of predicting total solids. The project positioned the University as a benchmark in technologies and process improvement, as well as bringing the company closer to data science and helping it make agile and informed decisions by reducing operational times in recipe updates.

Keywords: University-Industry Collaboration, Artificial Intelligence, Total Solids in Raw Milk, Predictive Model.

1 Introducción

La Universidad Nacional de Rafaela (UNRaf) se creó a partir de la sanción de la Ley N° 27.062 en el año 2014. Su puesta en marcha formó parte de un proceso de expansión del sistema universitario a partir del cual se añadieron 19 universidades públicas en el país.

La institución define en su misión "constituirse en un espacio institucional que contribuye a fortalecer el sistema educativo, científico y tecnológico en todos sus niveles y a formar personas altamente calificadas y comprometidas, capaces de diseñar y conducir las estrategias del sector productivo y laboral, anticipar los desafíos de la gestión territorial y ambiental y consolidar una cultura de la cooperación, la igualdad y la responsabilidad pública". (Universidad Nacional de Rafaela, 2021)

Además, establece en su visión: "trabajar como una institución de referencia regional, nacional e internacional, tanto en términos de innovaciones tecnológicas, pedagógicas, sociales y ambientales, como de vinculaciones institucionales y comunitarias, apoyada en el carácter federal de la configuración universitaria y los aportes de la Universidad Pública al desarrollo y la igualdad social". (Universidad Nacional de Rafaela, 2021)

Tomando como base estas definiciones, se establecieron y redactaron objetivos bajo cuatro perspectivas: desarrollo, gestión institucional, de articulación y científico académica. Esta última plantea un objetivo acorde a los fines de este trabajo que establece lo siguiente: "desarrollar la investigación y los procesos de transferencia y aplicación de conocimiento que contribuyan al avance científico, tecnológico y social fomentando la innovación en el territorio." (Universidad Nacional de Rafaela, 2021).

La Secretaría de Investigación y Transferencia Tecnológica de la universidad tiene como misión promover la investigación científica pertinente y de calidad. Está conformada por el centro de investigación aplicada UNRaf Tec y el área de gestión y administración denominada Unidad de Vinculación Tecnológica (UVT).[1]

En el año 2018 comenzó el funcionamiento del UNRaf Tec. En su reglamento se establece que "la investigación en el Centro actúa como enlace con las actividades

[1] La definición de la figura Unidad de Vinculación se encuentra estipulada en la Ley N°23.877 (Ley de promoción y fomento de la innovación tecnológica, 1990).

académicas y de formación de la UNRaf. La experiencia de investigación y de trabajo se volcará directamente en el trabajo cotidiano y en los cursos de las carreras de grado." (Universidad Nacional de Rafaela, 2024).

Tomando las definiciones presentadas hasta aquí, se puede afirmar que la UNRaf se constituye desde su creación como un espacio de articulación con las organizaciones a través de la ciencia y la tecnología.

El UNRaf Tec está constituido por Laboratorios de diversas disciplinas vinculados a un área temática de especialidad. Cada uno de ellos cuenta con su respectivo director/a y un equipo de gestión general.

En este trabajo se hace foco en un proyecto desarrollado por el Laboratorio de Gestión de la Información (LabGi). Este espacio aborda proyectos en temáticas como planificación estratégica; gestión basada en procesos; transformación digital; planes comerciales y de marketing; cultura basada en datos; inteligencia estratégica; vigilancia tecnológica e inteligencia competitiva.

En el mes de julio del año 2022, desde el LabGi, se generó un primer encuentro con una empresa láctea de la ciudad de Rafaela. En él, se encontraban cinco referentes de áreas clave que tenían un desafío común: incorporar inteligencia artificial (IA) para la mejora de procesos. Por parte de la UNRaf participó el Director de la Licenciatura en Administración y Gestión de la Información, la Directora del laboratorio y el especialista técnico en IA. Durante la jornada, se plantearon tres problemáticas a abordar por la empresa: estimación del ingreso de leche, estimación de la cantidad de sólidos en leche y optimización de recolecciones. La primera apuntaba a predecir el ingreso de leche al proceso productivo. En ese entonces, se realizaba una estimación a 12 meses y luego se ajustaba cada 3. La segunda tenía que ver con la posibilidad de predecir la cantidad de sólidos totales en la leche cruda. Este valor es crucial para determinar la calidad de la misma. La última proponía optimizar recorridos en base a rutas y recolecciones, sin afectar la calidad de la materia prima. El equipo de UNRaf explicitó que, para poder avanzar, era necesario que la empresa pudiera definir por cuál de las tres opciones continuar, considerando la disponibilidad de datos y el acompañamiento de un equipo de trabajo.

La compañía optó por el desafío 2: estimación de sólidos totales en la leche cruda.

La concentración de sólidos en la leche es uno de los indicadores para evaluar la calidad del producto y asegurar su valor nutritivo y aceptación en el mercado. Estudios recientes destacan cómo la variación en estos sólidos afecta la consistencia y la viscosidad, elementos esenciales para la aceptación del consumidor. Enriquez et al. (2012) demostraron que una mayor concentración de sólidos incrementa la consistencia y la viscosidad del yogurt. Por otra parte, se ha demostrado que variables como el índice de temperatura y humedad (ITH), las prácticas de alimentación y la genética del animal tienen un impacto directo en la composición de la leche. Dado el contexto de cambio climático y la variabilidad climática que afecta a las áreas de producción, el seguimiento y análisis de estos sólidos en zonas como Santa Fe, el estudio adquiere un nuevo nivel de relevancia para la empresa. Asimismo, otro aspecto fundamental por el cual una correcta estimación de sólidos representa una importancia para las empresas lácteas, es

que dicha variable constituye un insumo de entrada para su proceso productivo, modificando la composición de los productos y el tratamiento que adquiere la materia prima para la elaboración de los mismos.

El estudio propone desarrollar un modelo predictivo que permita optimizar la calidad del producto final. El resultado puede utilizarse como herramienta estratégica para la industria, facilitando ajustes en el manejo y prácticas de alimentación según las proyecciones de calidad, asegurando una producción lechera eficiente.

Para el tratamiento de problemas similares donde se dispone de información histórica existen distintos enfoques factibles de ser aplicados, siendo los métodos de aprendizaje automático uno de los más prometedores tal como se detalla en estudios previos (Ji et al., 2022; Hansen et al., 2024).

Para el desarrollo del trabajo, la empresa definió el equipo responsable. En los primeros encuentros se avanzó en profundizar técnicamente la problemática a abordar. Se conversó sobre supuestos por parte del equipo, descripción del proceso actual y determinación de variables que afectan la variación de los sólidos totales en la leche cruda.

Finalmente, se planteó un cronograma de actividades basado en cuatro etapas y con una proyección de seis meses:

1) Recopilación y validación de datos pertinentes
2) Elaboración de hipótesis y análisis exploratorios de datos
3) Codificación, ejecución y validación de modelos de estimación
4) Desarrollo de entregables técnicos y documentales

2 Metodología

Se trata de una investigación cuantitativa (Hernández Sampieri et al., 2014) puesto que se siguió un proceso estricto compuesto de una serie de pasos, donde se realizó una delimitación del problema, se formularon hipótesis, se recolectaron los datos pertinentes y luego los mismos fueron analizados mediante métodos estadísticos para arribar a una conclusión sobre dichas hipótesis.

A partir de esto se seleccionó la metodología CRISP-DM (Shearer, 2000), la cual se detalla en la **Fig. 1**.

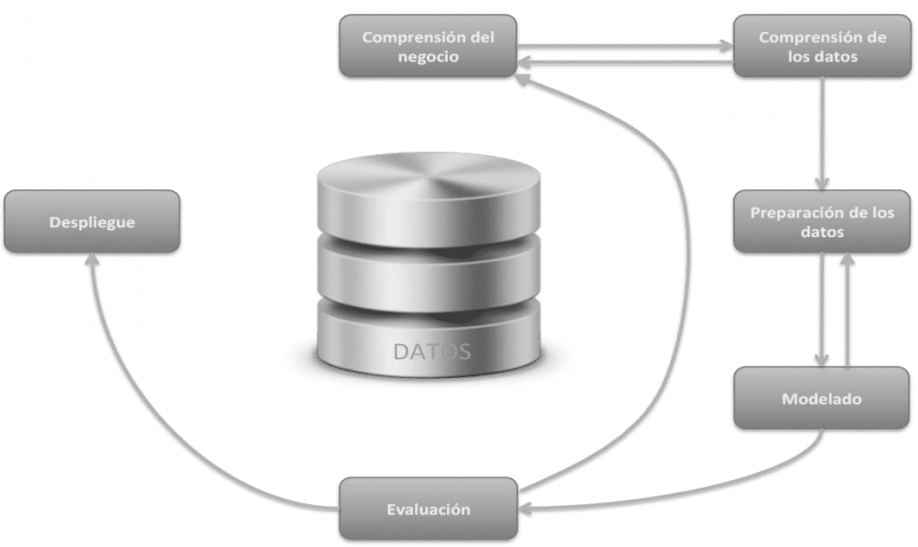

Fig 1. Detalle de las etapas y secuenciación de las mismas planteadas en la metodología CRISP-DM.

Como parte del entendimiento del negocio, se realizaron numerosas reuniones con el equipo técnico de la compañía y el equipo de trabajo donde se comenzó inicialmente con una descripción completa del proceso para comprender cuales son las variables que afectan a la problemática de interés, y cómo impactan las mismas dentro del proceso productivo. Este aspecto fue clave por dos factores: para comprender cómo modelar el problema desde un punto de vista técnico y cómo debería ser el entregable a realizar, de forma que pueda ser utilizado por los/as usuarios/as de una forma conveniente (pudiendo contribuir positivamente a la operativa y sin dañar los procesos existentes). El conocimiento brindado por el equipo de trabajo fue respaldado por un proceso de revisión bibliográfico para reafirmar e introducir propuestas que fueron desarrolladas previamente (las cuales se describieron en la sección 1 de este artículo).

Como parte de la segunda etapa, se llevó a cabo un relevamiento detallado de la información disponible. La misma se encontraba en las condiciones apropiadas de ser utilizada para esta iniciativa puesto que era consumida por distintas áreas de la empresa para la generación de informes. Esto significó una ventaja respecto a la exportación de los datos y la posibilidad de establecer relaciones entre las distintas fuentes consultadas. Una vez realizada la extracción desde las distintas áreas encargadas de generar dicha información, se realizaron distintos análisis exploratorios de datos con los siguientes objetivos:

- Realizar un entendimiento de los mismos de parte del equipo técnico
- Validar que el comportamiento de ciertas variables mencionadas en las reuniones previas sea el esperado
- Presentar hallazgos o situaciones particulares que no habían sido advertidas o mencionadas previamente por la empresa, que pudieran tener relevancia en esta iniciativa o en otras futuras

La preparación de los datos consistió en la determinación de las variables a ser utilizadas y su correspondiente procesamiento para ser aplicado en los modelos estadísticos. La detección y procesamiento de valores nulos se abordaron en esta etapa también.

La instancia de modelado incluyó la definición de un esquema de experimentación que permitió llevar adelante pruebas con distintos algoritmos de aprendizaje automático, realizando una separación en entrenamiento y prueba (mediante validación cruzada). Esto sirvió para analizar las capacidades de generalización de los modelos utilizados. La variable a predecir fue el porcentaje de sólidos promedio mensual del mes próximo, considerando todas las entregas realizadas en planta. Esto quiere decir que los datos de entrada utilizados son calculados hasta el mes **t**, y la variable a predecir refleja el valor del mes **t+1**.

Para realizar la evaluación de cada experimento se definió como métrica el error promedio, que debido a que la variable representa un porcentaje en sí mismo, puede ser equivalente al error porcentual promedio tal como se describe en (**1)** expresado sobre un total de *n* muestras, donde para cada muestra *i* se tiene el valor predicho *y* y el valor real $\hat{y}$. Esta variable se seleccionó debido a que es útil desde el punto de vista técnico y permite ponderar los errores realizados por los modelos de manera apropiada, y al mismo tiempo presenta una fácil interpretación para usuarios no técnicos.

$$ECM = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (1)$$

Ecuación 1. Error cuadrático medio

Como herramientas durante la fase de experimentación se utilizaron librerías de código abierto desarrolladas en Python, principalmente pandas 2.1 [2] para el manejo de datos, scikit-learn [3] y statsmodels [4] para la creación de los modelos utilizados. A su vez, la librería scikit-learn fue utilizada también para realizar la separación de los datos y una búsqueda en grilla de hiper-parámetros en cada modelo. Los valores de hiper-parámetros explorados así como los resultantes luego de cada adaptación no han sido incluidos en el informe por cuestiones de confidencialidad por parte de la empresa.

Luego de realizar numerosas iteraciones de las etapas detalladas previamente, se arribó a la selección de un modelo que presentaba el menor error en la métrica seleccionada utilizando un conjunto de variables específicas. A partir de dicho modelo se llevó a cabo el desarrollo de un código en el lenguaje de programación Python capaz de procesar una serie de archivos generados por un/a usuario/a, extraer las variables de interés, entrenar el modelo de aprendizaje automático y generar las predicciones. Como soporte a la ejecución de este proceso se utilizó la herramienta Google Colab [5], debido a que la empresa contaba con la utilización de la GSuite como paquete comparativo.

[2] https://pypi.org/project/pandas/2.1.0/

[3] https://pypi.org/project/scikit-learn/1.2.0/

[4] https://pypi.org/project/statsmodels/0.14.0/

[5] https://colab.google/

3 Resultados

3.1 Resultados técnicos

Durante la fase de experimentación se llevaron a cabo distintas pruebas utilizando diferentes algoritmos (entre ellos: regresión lineal, árboles de decisión, k-vecinos más cercanos, redes perceptrón multicapa - MLP - y ARIMA). Las variables utilizadas fueron las siguientes:

- precipitaciones promedio
- humedad promedio
- temperatura promedio
- estación
- trimestre
- sólidos totales actuales
- proteínas actuales
- materia grasa actual
- suma de litros actual

En todos los casos se evaluaron diferentes combinaciones en búsqueda del mejor conjunto de variables de entrada para ser utilizado. En cada modelo también se llevó a cabo una búsqueda en grilla para optimizar la combinación de hiper-parámetros en los distintos modelos. La **Tabla 1** introduce los resultados promedio alcanzados por la mejor combinación de cada uno de los distintos modelos en el conjunto de entrenamiento y en el de validación.

Tabla 1. Error cuadrático medio de los distintos modelos analizados en los conjuntos de entrenamiento y validación.

Modelo	Entrenamiento	Validación
Modelo actual	0,289	0,219
Regresión lineal	0,056	0,105
Árboles de decisión	0,096	0,297
k-vecinos más cercanos	0,085	0,289
Perceptrón multicapa	0,085	0,138
ARIMA	0,091	0,257

En base a los resultados presentados, se puede resaltar que la regresión lineal es el modelo que obtuvo el mejor rendimiento en ambos conjuntos de datos. Debido a que el conjunto de datos utilizado presentaba un número de filas limitado, los resultados parecen indicar que dicho modelo presentaba un balance apropiado entre la cantidad de

parámetros necesarios y el volumen de información disponible para ser entrenado (en comparación con otros modelos tales como el perceptrón multicapa por ejemplo).

Finalmente, en la **Tabla 1** se introducen los resultados del enfoque utilizado por la empresa pudiendo notar que mediante las otras propuestas existen mejoras evidentes respecto a la disminución en la métrica evaluada.

3.2 Resultados de vinculación y lecciones aprendidas

Tal como se detalló al inicio de este trabajo, existe un gran interés por parte de la Universidad Nacional de Rafaela respecto de la vinculación con el sector productivo, entendiendo a éste como parte fundamental dentro del rol que ocupa la institución con la sociedad. En este sentido, la colaboración ha beneficiado a ambas partes. La empresa ha accedido a conocimientos y tecnologías de vanguardia, mejorando sus procesos y fortaleciendo su competitividad. Asimismo, logró un primer acercamiento a la ciencia de datos participando de este tipo de iniciativas y pudiendo experimentar de manera cercana las distinta etapas que involucran a proyectos de este tipo. Como aspecto destacable de este punto se señala la oportunidad de generación de nuevas iniciativas producto de una mayor comprensión respecto al uso del dato como insumo fundamental para la toma de decisiones.

La casa de estudios, por su parte, ha tenido la oportunidad de incorporar a una estudiante al proceso y de aplicar sus capacidades en un contexto real, generando un impacto tangible en el sector productivo y fortaleciendo su rol como agente de desarrollo regional.

Esta experiencia sienta las bases para futuras colaboraciones y abre nuevas oportunidades para la transferencia de tecnología y conocimientos. Como lección aprendida, proponer un esquema escalable de colaboración universidad-industria que puede ser replicable en otras instancias. En lo que respecta al concepto de escalabilidad, se propone llevar a cabo acciones concretas y con períodos de entre 2 y 4 meses que permitan evidenciar resultados tangibles y confianza entre las partes. En experiencias similares con otras empresas, se propone una acción inicial de formación / transferencia de conocimiento por parte de la Universidad, luego una asistencia técnica y, de continuar, desarrollo de propuestas formativas y/o trabajos de transferencia de mayor duración.

4 Conclusiones

En términos de vinculación, el trabajo presentado en este documento fue el iniciador de una relación entre una reconocida empresa de la localidad de Rafaela, Santa Fe y una universidad joven de la misma ciudad: la UNRaf.

En cuanto a los aspectos técnicos, la transferencia tecnológica permitió proponer una mejora a un proceso clave para la empresa. Se logró desarrollar una herramienta automática de estimación trimestral de sólidos totales, grasas y proteínas. A partir de esta, se acercó al equipo de trabajo de la empresa a la ciencia de datos y a la generación de modelos predictivos. Estas tecnologías de carácter emergente son parte de la génesis de los proyectos que desarrolla UNRaf: la mirada tecnológica como eje central de la mejora.

La mejora del proceso tuvo un impacto de carácter operativo y estratégico: reducir los tiempos de actualización de recetas e incentivar la toma de decisiones basada en datos.

La experiencia desarrollada ha abierto nuevas oportunidades de vinculación entre las partes, dado que, en 2025, la empresa se acercó nuevamente con otro desafío de similares características.

Referencias

Universidad Nacional de Rafaela (2021). Estatuto definitivo de la Universidad Nacional de Rafaela

Universidad Nacional de Rafaela (2024). Ordenanza Nº 004/2024. Reglamento del centro de investigación y transferencia tecnológica UNRaf-TEC de la Universidad Nacional de Rafaela.

Enriquez, D., Sánchez-González, J., & Castro Santander, P. (2012). Efecto de la concentración de sólidos totales de la leche entera y tipo de cultivo comercial en las características reológicas del yogurt natural tipo batido. Agroindustrial Science, 2, 173-180.

Ji, B., Banhazi, T., Phillips, C. J. C., Wang, C., Li, B. (2022). A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. Biosystems Engineering, vol 216, 186-197.

Hansen, B. G., Li, Y., Sun, R., & Schei, I. (2024). Forecasting milk delivery to dairy – How modern statistical and machine learning methods can contribute. Expert Systems with Applications, 248, 123475.

Thomas, J., Ramos, E., Gieco, J., Jáuregui, J., Badino, O., Leva, P., & Toffoli, G. (2014). Evolución de la concentración de sólidos útiles en leche de tambos del NE de la Provincia de Santa Fe: Período 2003-2013. Revista FAVE - Ciencias Agrarias, 13 (1–2), 1–14.