



Tecnológico de Monterrey

Campus Querétaro

Reporte Final

Gamaliel Marines Olvera	A01708746
Uri Jared Gopar Morales	A01709413
José Antonio Miranda Baños	A01611795
María Fernanda Moreno Gómez	A01708653
Oskar Adolfo Villa López	A01275287
Luis Ángel Cruz García	A01736345

Inteligencia artificial avanzada para la ciencia de datos II
Grupo 501

Reporte Final

Introducción

En este documento se presentan los resultados obtenidos a lo largo del proceso de análisis y desarrollo de un modelo de clasificación de imágenes basado en redes neuronales convolucionales (CNN). El modelo está diseñado para clasificar imágenes en tres categorías: "cama vacía", "vaca acostada" y "vaca de pie".

El objetivo principal de este documento es explicar de manera clara y concisa los resultados obtenidos y su implementación, generando información valiosa que apoye la toma de decisiones sobre el uso de las camas y los periodos en los que se registran mayores niveles de ocupación.

En las siguientes secciones, se detallarán los beneficios identificados, el proceso seguido, los costos asociados, las dificultades que desviaron el proyecto de su plan original, y los planes de implementación, como manuales y guías. Estos elementos permitirán al usuario replicar el proyecto con facilidad. Asimismo, se incluirán recomendaciones basadas en la experiencia obtenida durante el desarrollo.

Resultados Obtenidos

Tenemos dos tipos de resultados en este proyecto. **Resultados del Modelo** que evalúan principalmente el desempeño del modelo, para asegurarse de que el modelo se ajusta al problema y puede capturar de forma correcta las diferencias entre las clases. **Resultados de Análisis de Datos** nos indican hallazgos que se hayan encontrado mediante el modelo, estos hallazgos son de especial interés, puesto que permitirán tomar decisiones sobre la industria. Para ambos casos conseguimos resultados positivos, cumpliendo así con nuestros objetivos.

- **Resultados de Modelado:** Después de varias iteraciones de modelado conseguimos un modelo con las siguientes métricas, un F1 score de 96%, un accuracy de 96% y un accuracy promedio por cama de 95%, el cual fue el más alto de nuestras iteraciones. Estas métricas nos indican que el modelo distingue de forma correcta entre clases en distintos casos.
- **Resultados de Análisis de Datos:** Con el modelo ajustado y usando el dataset completo obtuvimos los siguientes hallazgos, la cama C es la cama menos utilizada con 9.5% de utilización, la cama A tiene una utilización del 43.7% y la cama B siendo la más usada con un 46.8% de utilización. Esto nos indica que la cama C es evitada por las vacas en general. Por último encontramos un patrón de uso en las camas, siendo ocupadas con mayor frecuencia en la madrugada y en la noche.

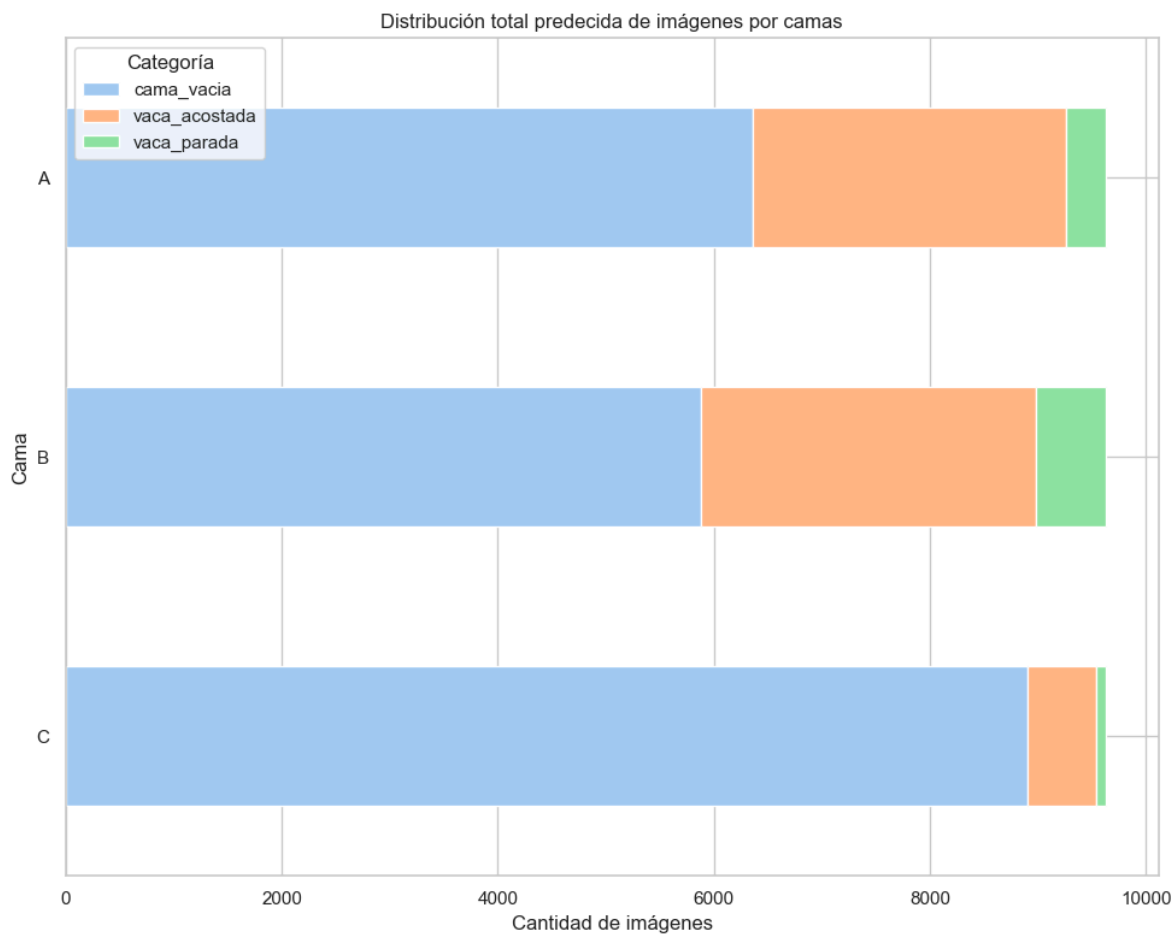
Hallazgos:

Imagen 1: Distribución de las imágenes de cada cama de acuerdo a las categorías. Dataset completo. Predicción del modelo.

Se analizó cuál era la cama con más uso. De todas las imágenes etiquetadas como vaca acostada por el modelo, se graficó la distribución por cama en porcentaje. Aquí confirmamos que la cama C es la menos usada, solo en un 9.5 % de las veces. Las camas A y B no tienen una diferencia significativa en su uso.

Distribución de vacas acostadas por cama

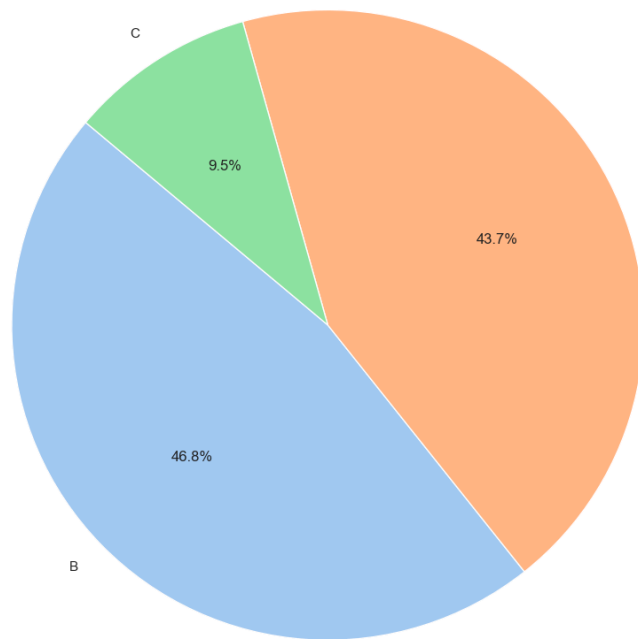


Imagen 2: Distribución de las imágenes de cada cama de acuerdo a las categorías. Clasificación manual.

Con la información obtenida, también es posible encontrar un histograma de las vacas acostadas en cada cama distribuido en cada hora del día, así como un histograma de las vacas acostadas en todas las camas combinadas. Esta información es útil para saber en qué momentos del día prefieren acostarse las vacas.

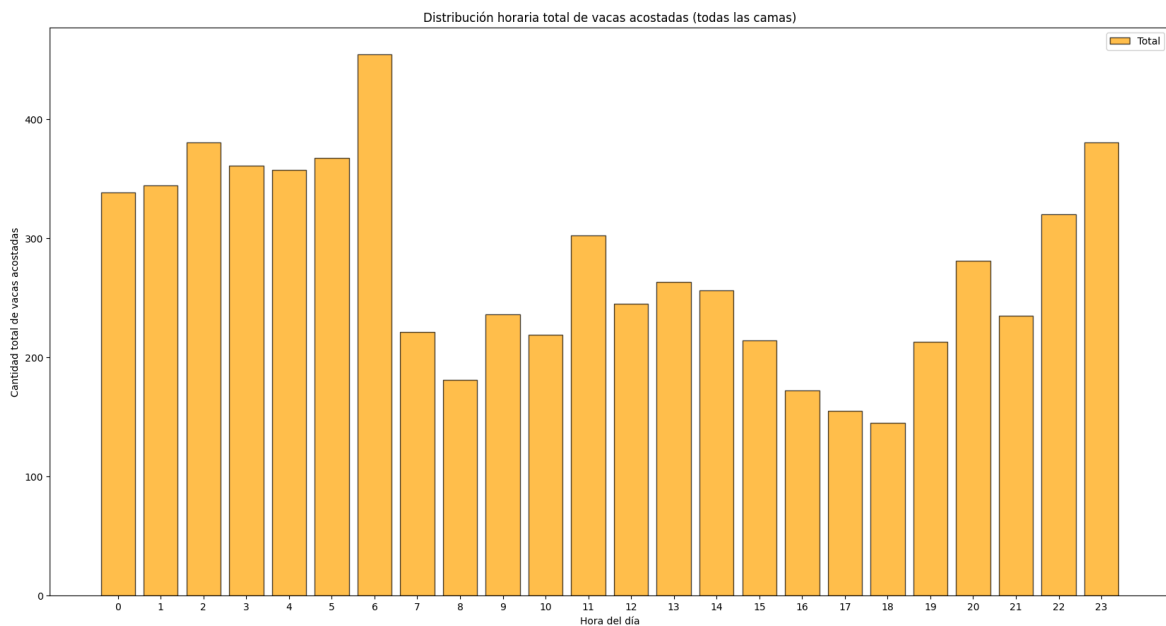


Imagen 3: Observación de horas en las que las vacas ocupan más las camas

La **imagen 3** es la distribución general, nos indica que en general las camas son ocupadas en la madrugada y en la noche, mientras que presentan un uso menor de las 7 de mañana hasta las 6 de la tarde.

Procesos y Costos

Siguiendo la metodología de CRISP DM tuvimos varios procesos con diversas formas de costo (Monetario, Tiempo, y Personal), estos son descritos en la tabla siguiente.

Nombre	Descripción	Costo
Business Understanding	Entender la problemática y el negocio. Plantear un objetivo de análisis.	Monetario: \$0,00 Tiempo: 7.88 h Personal: 6 personas
Data Understanding	Revisar los datos dados brindados, entender su composición y tipo.	Monetario: \$0,00 Tiempo: 10.1 h Personal: 6 personas
Data Preparation	Con base en Data Understanding modificar el conjunto de datos con el fin de poder modelarlos correctamente.	Monetario: \$0,00 Tiempo: 4.2 h Personal: 8 personas

Modeling	Generar modelos mediante varias iteraciones que nos permitan generar conocimiento nuevo y de utilidad.	Monetario: \$0,00 Tiempo: 35.5 h Personal: 4 personas
Evaluation	Se verifica si el modelo cumple los objetivos del negocio, si los hallazgos son consistentes con las metas y se revisa el proceso para identificar pendientes y guiar actividades futuras	Monetario: \$0,00 Tiempo: 35.5 h Personal: 4 personas
Deployment	La fase final integra el modelo en producción o entrega resultados mediante informes, paneles o API y este reporte.	Monetario: \$0,00 Tiempo: 8 h Personal: 3 personas

Desviaciones del plan original

A lo largo del proyecto tuvimos desviaciones en diferentes fases, mayormente en la parte de **Modelado**, esto fue ocasionado principalmente por no haber definido métricas funcionales desde el principio, lo que pudieron ser dos iteraciones de modelado, o una, se convirtieron en tres. El principal problema durante el **Modelado** fue la diferencia entre las camas, que creímos habíamos corregido durante la etapa de **Data Preparation** cuando alargamos las imágenes para poder aprovechar toda el área de la cama. Esta diferencia resultó en modelos que se especializaban en las camas más concurridas e ignoraban las camas que se usaban con menor frecuencia. Esto en el análisis se reflejaba en datos relativamente correctos en la cama A y B, sin embargo, terminamos teniendo predicciones falsas, que no seguían las tendencias de las muestras iniciales en la cama C. La solución fue el modelo tres que hace uso de una aumentación que involucra el brillo para ajustarse mejor a la cama C, junto a esto desarrollamos la métrica de Accuracy promedio de camas que nos reflejaba el promedio del accuracy individual de cada cama, lo que nos permitió comparar los modelos de forma más justa para todas las camas.

Otro problema que enfrentamos en la parte de **Modelado** fue al tratar de hacer pruebas de arquitectura, es decir, probar nuestro programa dentro de hardware igual al que se nos especificó que estaba dirigido el proyecto, ya que no disponíamos del hardware exacto y decidimos reemplazarlo con una versión anterior del mismo,

pensando que “*Si funciona en el peor hardware, no tendrá problema trabajando en un hardware moderno*”. Esto nos costó alrededor de una semana y terminamos optando por adquirir el hardware especificado en un inicio.

Planes de Implementación

En este apartado se redactará sobre las acciones necesarias para poner en marcha los resultados del modelo y lograr la correcta implementación de nuestro modelo con la arquitectura que nos fue especificado.

Integración de la solución al CAETEC:

Para la implementación de nuestra solución en la arquitectura del CAETEC, se proporcionará una carpeta de Google Drive, la cual cuenta con nuestra interfaz de usuario y pesos entrenados del modelo, esto para poder instalar toda la paquetería necesaria para correr nuestro modelo, además, cuenta con las funciones necesarias para cargar el modelo y de esta manera se pueda implementar en el cualquier equipo de 64 bits.

Despliegue del modelo:

Para el despliegue del modelo y la instalación del proyecto en general, se redactó un manual de usuario en el que se detallan las fases a seguir, se adjunta el enlace a continuación:

https://github.com/Mirabay/Vakas/blob/main/docs/manuals/Technical_User_Manual_TC.pdf

Recomendaciones de Trabajo Futuro

En esta sección colocamos una lista de recomendaciones a futuro, pensando en que este proyecto ha dejado varias lecciones entre sus participantes.

- Generar métricas basándonos en data understanding y data preparation.
- Hacer pruebas estadísticas con la muestra inicial brindada.
- Ocupar lo menos posible el Accuracy tradicional.
- Contar con el hardware adecuado en todo momento, en especial en la prueba de arquitectura.