Análisis de proyectos basados en Minería de datos y CRISP-DM

Pedro González Fernández

29 de octubre de 2025

Índice

1. TFG Estimación de edad a partir de la rodilla 2

1.1. Comprensión del negocio 2
1.2. Comprensión de los datos 2
1.3. Preparación de los datos 2
1.4. Modelado 3
1.5. Evaluación 3
1.6. Despliegue o aplicación final 3

2. Análisis de tráfico y optimización de rutas con machine learning 4

2.1. Comprensión del negocio 4
2.2. Comprensión de los datos 4
2.3. Preparación de los datos 4
2.4. Modelado 4
2.5. Evaluación 4
2.6. Despliegue o aplicación final 4

3. Comparación de ambos trabajos 5

3.1. ¿Qué proyecto representa mejor las fases del CRISP-DM? 5
3.2. ¿Qué tipo de modelos o técnicas se aplican en cada caso? 5
3.3. ¿Qué aspectos podrías aplicar tú en un proyecto propio de minería de datos? 5

4. Bibliografía 5

1. TFG Estimación de edad a partir de la rodilla

El primer proyecto que he encontrado y voy a analizar trata sobre el **desarrollo de un sistema basado en deep learning para la estimación de la edad usando la imagen radiolñogica de rodilla**, de la UGR.

1.1. Comprensión del negocio

Este proyecto está realacionado con el ámbito de la **Antropología Forense**, donde la estimación de la edad es un proceso esencial, tanto a la hora de identificar restos humanos, como en casos legales que involucran personas vivas sin documentación. El autor parte de la necesidad social y científica de desarrollar métodos más precisos, rápidos y objetivos, empleando la IA como apoyo al trabajo de forense.

El problema se centra en la dificultad de estimar la edad mediante la rodilla, –una parte del cuerpo poco exlotada a nivel anatómico– y el objetivo es crear un sistema capaz de realizar esta tarea de forma automática mediante imágenes de resonancia magnética.

El autor incluye en esta fase la relevancia ética y práctica del modelo, aludiendo a su posible aplicación en migración, identificación y procesos judiciales, dónde es crucial la precisión de la estimación.

1.2. Comprensión de los datos

El proyecto trabaja con imágenes médicas tridimensionales de resonancia magnética nuclear. En esta fase se analiza el tipo de datos disponibles, su formato, su distribución por edades y las característiacas anatómicas visibles en las imágenes. Se examina la calidad de las imágenes y se toman decisiones sobre si es necesario aplicar técnicas de mejora visual y reducción de ruido. También se tienen en cuenta los metadatos clínicos que contienen las imágenes, que posteriormente se integran como variables complementarias en el modelado.

La exploración inicial permite detectar las limitaciones en la cantidad de datos y su rango de edad, aspectos que condicionan el rendimiento y generalización de los modelos.

1.3. Preparación de los datos

Esta fase ocupa una gran parte del proyecto. Implementa preprocesamiento de imágenes para mejorar la calidad de la información visual que se usará en el entrenamiento. Emplea técnicas como la normalización de intensidades, ajuste de contraste y, sobre todo, el método de umbralización de Otsu para segmentar las estructuras relevantes de la rodilla. También se describe la generación de varios subconjuntos de datos para el entrenamiento, la validación y la prueba del modelo, con especial cuidado en la representatividad de las edades en cada uno. De forma adicional, integra metadatos (como el sexo) en algunos modelos para evaluar si mejoran la precisión del sistema. En esta fase se transforma un conjunto de imágenes médicas heterogéneas en un dataset estructurado y listo para el modelado.

1.4. Modelado

Se desarrollan y entrenan varios modelos de deep learning, centrándose en arquitecturas como ResNet y DenseNet. También se comparan dos enfoques conceptuales diferentes:

- Uno basado en regresión, para estimar directamente la edad numérica.
- Otro de clasificación, para asignar rangos de edades o distinguir entre menores y adultos.

Los modelos emplean PyTorch, –libreria open-source usada para deep learning y procesamiento de lenguaje natural– y se ejecutan en un servidor proporcionado por el Instituto DaSCI. esta fase representa el núcleo experimental del proyecto, donde se empieza a materializar la solución técnica al problema definido.

1.5. Evaluación

Los modelos que se han desarrollado se someten a una evaluación exhasutiva mediante métricas como el **MAE (Mean Absolute Error)**. Se comparan las diferentes variantes y arquitecturas y se analizan los resultados en profundidad, identificando las configuraciones que ofrecen mejores resultados. Además, se comparan los resultados obtenidos con los de otros estudios, valorando avances y limitaciones del modelo en relación con los requisitos de precisión exigidos para su eso forense real. Esta fase termina en un análisis crítico de la fiabilidad y aplicabilidad del sistema.

1.6. Despliegue o aplicación final

Pese a que el trabajo tiene una orientación investigadora, se plantea un sistema funcional capaz de estimar la edad a partir de una nueva imagen de resonancia magnética de rodilla, sin necesidad de intervención humana. Se documenta la implementación del código en un repositorio público de GitHub y se describe la arquitectura modular del sistema, lo que permite su futura integración en entornos reales. En las conclusiones se señalan los pasos necesarios para una aplicación práctica, mencionando la mejora del conjunto de datos y la validación de nuevas poblaciones. De este modo, el despliegue es una base sólida sobre la que construir una herramienta real de apoyo pericial.

2. Análisis de tráfico y optimización de rutas con machine learning

El segundo proyecto trata sobre un modelo para **analizar y optimizar el tráfico mediante deep learning**, de la UOC.

2.1. Comprensión del negocio

En esta fase se definen los principales objetivos y la problemática a resolver del proyecto. Se establece la necesidad de analizar los datos de tráfico de Madrid para optimizar las rutas mediante modelos de machine learning, identificando el propósito, alcance y beneficios esperados.

2.2. Comprensión de los datos

Se revisa de forma exhaustiva las fuentes de información disponibles, y se selecciona el conjunto de datos abierto del Ayuntamiento de Madrid, por su fiabilidad, formato estándar y volumen más que suficiente para el análisis. En esta etapa se analizan las características de las variables, su estructura y la calidad de los registros, identificando posibles valores nulos o que no son válidos.

2.3. Preparación de los datos

Una vez comprendidos los datos, se llevan a cabo labores de integración, limpieza y transformación. Se fusionan ficheros mensuales, se añaden nuevas columnas como el día de la semana, y se ajustan los valores de intensidad y velocidad media. En esta fase se acaba generando un conjunto de datos homogéneo y listo para modelar.

2.4. Modelado

En esta fase se aplican algoritmos y modelos de aprendizaje automático de optimización de rutas. Se desarrollan y comparan enfoques como el algoritmo de Dijkstra, la colonia de hormigas (ACO) o la colonia de abejas (ABC), empleando parámetros de tráfico y velocidad para rutas mas eficientes. Cada modelo se configura y ejecuta sobre los datos procesados para evaluar su rendimiento y viabilidad.

2.5. Evaluación

Se analizan los resultados obtenidos por los modelos, comparando la efectividad de cada uno según criterios como la ditancia y duración de trayecto o el tiempo de ejcución. Se verifica que cumplan con los objetivo iniciales y se evalua si es posible su aplicación en entornos reales.

2.6. Despliegue o aplicación final

Finalmente, se integran los resultados en una propuesta funcional que demuestra la viabilidad de aplicar sistemas Big Data y algoritmos de IA para la gestión

del tráfico. Se elaboran mapas de calor, visualizaciones interactivas y ejemplos de optimización de rutas, estableciendo la base para futuras mejoras, como incorporar datos en tiempo real o ampliar el área de estudio.

3. Comparación de ambos trabajos

3.1. ¿Qué proyecto representa mejor las fases del CRISP-DM?

Los dos trabajos cubren las fases del modelo CRISP-DM, aunque se centran en distintas etapas. Mientras que el primero se centra y detalla las fases de **Modelado** y **Evaluación**, el segundo tiene una implementación muy clara de las fases de **Comprensión y preparación de los datos**.

3.2. ¿Qué tipo de modelos o técnicas se aplican en cada caso?

El primero se centra en el deep learning, empleando técnicas como redes neuronales convolucionales, usando arquitectura ResNet para procesar las imágenes.

El segundo trata de optimización, usando algoritmos como Dijkstra o la colonia de hormigas. Además, también emplea redes neuronales convolucionales.

3.3. ¿Qué aspectos podrías aplicar tú en un proyecto propio de minería de datos?

Sobretodo aspectos como no limitarme a probar un modelo estándar, sino investigar y experimentar con varios; integrar librerias específicas que me permitan modelar datos complejos; definir y medir métricas con sentido en el contexto del problema, etc.

4. Bibliografía

Estimación de edad a partir de la rodilla:

https://digibug.ugr.es/handle/10481/105807

Análisis de tráfico y optimización de rutas con machine learning:

https://openaccess.uoc.edu/items/330072f2-7827-4268-9ef3-a4f32b8d04b0#page=1