Plan de Pruebas

Este plan de pruebas describe las actividades diseñadas para evaluar la calidad y el rendimiento del sistema de estimación ósea pediátrica a partir de carpogramas, desarrollado en el marco de este proyecto de grado. El objetivo principal es validar si el sistema cumple con los estándares mínimos de precisión definidos a partir del estado del arte, y asegurar que su comportamiento sea robusto y consistente bajo diferentes configuraciones de entrenamiento y datos.

Objeto a evaluar

El objeto de evaluación corresponde al sistema de inteligencia artificial desarrollado para predecir la edad ósea pediátrica a partir de imágenes de carpograma. Específicamente, se evaluarán el modelo seleccionado verificando su desempeño y robustez.

Métricas y criterios de aceptación

Para medir el rendimiento del sistema se utilizarán dos métricas estándar en problemas de regresión:

- R^2 (Coeficiente de Determinación): Evalúa qué proporción de la variabilidad de la variable dependiente puede ser explicada por el modelo. Un valor cercano a 1 indica alto poder explicativo.

- MAE (Error Absoluto Medio): Mide el promedio de los errores absolutos entre la predicción y el valor real. Se expresa en años para facilitar la interpretación clínica.

Cabe aclarar que, se evaluarán solamente las métricas de la fase de validación contra los criterios de éxito definidos a continuación, esto con el fin de hacer una verificación realista de la calidad del modelo.

Umbrales de aceptación

Los criterios de éxito se definieron con base en el estudio de Dehghani et al. (2019), quienes emplearon Support Vector Regression sobre 442 radiografías de individuos de 0 a 18 años, obteniendo un R^2 de 0.738 y un MAE de 0.55 ± 0.1 años. Este estudio fue seleccionado como punto de referencia por su similitud en tamaño de muestra y enfoque metodológico.

Criterios mínimos para considerar un modelo aceptable:

- $R^2 \geq 0.738$

- MAE ≤ 0.65 años (equivalente a 237.25 días)

Instrumentos y entorno de prueba

- Herramientas utilizadas: Jupyter Notebook, Python 3.6.13, bibliotecas scikit-learn 0.24.2, xgboost 1.5.2, tensorflow 2.6.2, keras 2.6.0, pandas 1.1.5 y numpy 1.19.5.

- Entradas del sistema: Características radiómicas estructuradas (.csv) combinadas con metadatos anonimizados.

- Salidas esperadas: Predicciones de edad ósea expresadas en días, convertidas posteriormente a años para análisis clínico.

Casos de prueba

Se definieron tres tipos de casos de prueba, aplicados a cada modelo y configuración de segmentación:

Código	**Descripción del caso**	**Conjunto de datos**	**Resultado esperado**
CP1	Evaluación del modelo general	Datos completos	Cumple umbrales de R^2 y MAE
CP2	Evaluación del modelo especializado masculino	Solo datos masculinos	Cumple umbrales de R^2 y MAE

CP3	Evaluación del modelo especializado femenino	Solo datos femeninos	Cumple umbrales de R^2 y MAE

Tabla 7. Casos de prueba

En todos los casos, se considerará el cumplimiento del criterio como satisfactorio si ambas métricas están dentro del rango especificado.

Ejecución de pruebas

La ejecución se realizó en entorno local, mediante scripts reproducibles que aplican la predicción sobre los conjuntos de prueba, validación y entrenamiento. Las métricas fueron calculadas con las funciones mean_absolute_error y r2_score de sklearn.metrics. El MAE fue convertido a años mediante la fórmula:
MAE (años) = MAE (días) / 365

Conclusiones esperadas de las pruebas

- Si un modelo supera ambos umbrales, se considera que su desempeño es clínicamente aceptable.

- Si uno o ambos criterios no se cumplen, el modelo será considerado no apto bajo los estándares establecidos.

Conclusiones de las pruebas

Con base en los umbrales definidos ($R^2 \geq 0.738$ y MAE ≤ 0.65 años), y tomando como referencia la fase de validación, se obtuvo el siguiente comportamiento para los modelos evaluados:

Modelo general (XGBoost con segmentación Otsu):

- **R^2:** 0.963
- **MAE:** 0.278 años (equivalente a 101.67 días)

Este modelo superó ampliamente ambos umbrales, siendo considerado clínicamente aceptable y el más robusto de todos los evaluados. Su equilibrio entre precisión y eficiencia lo consolidó como el modelo seleccionado para la solución final.

Modelo especializado por género masculino (XGBoost masculino con segmentación Otsu):

- **R^2:** 0.896
- **MAE:** 0.909 años

Aunque el R^2 supera el umbral de aceptación, el MAE excede el límite definido, lo cual sugiere una menor precisión clínica. Se considera un resultado aceptable en términos explicativos, pero no suficiente para justificar su adopción frente al modelo general.

Modelo especializado por género femenino (XGBoost femenino con segmentación Otsu):

- R^2: 0.747
- MAE: 1.187 años

Este modelo apenas alcanza el umbral mínimo de R^2, pero su MAE está considerablemente por encima del rango aceptable. Por tanto, a pesar de ser un resultado aceptable en términos explicativos, no es suficiente para justificar su adopción frente al modelo general.

En resumen, de los tres modelos evaluados, únicamente el modelo general con segmentación Otsu cumplió de forma rigurosa con todos los criterios establecidos, confirmando su idoneidad para su implementación en contexto clínico. Los modelos especializados por género, si bien arrojaron resultados interesantes, no superaron de forma consistente los umbrales definidos y no se consideraron superiores al modelo general en esta validación.