Services and Applications for a Healthy Ageing
Servicios y Aplicaciones para Envejecimiento Saludable

SAPIENS

Este proyecto ha sido concedido por la Agencia Estatal de Investigación del Ministerio de Ciencia e Innovación dentro del procedimiento de concesión de ayudas a 'Proyectos de I+D+i' en el marco de los programas estatales de generación de conocimiento y fortalecimiento científico y tecnológico del sistema de I+D+i y de I+D+i orientada a los retos de la sociedad, convocatoria 2020.

MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

Proyecto de investigación financiado por

MCIN/AEI/10.13039/501100011033

D.4.2: First computational model and analysis results report

WP4: Data Analytics

UNIVERSIDADE DE VIGO

atlanTTic research center for Telecommunication Technologies

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tabla de contenido

1. Introducción ***2***

2. Modelo computacional ***3***

1. Recopilación de Datos 3
2. Preprocesamiento de Datos 3
3. Selección de Características (Feature Selection) 3
4. Validación Cruzada y Optimización de Hiperparámetros 4

3. Resultados de la analítica de datos ***5***

4. Personal participante ***7***

1. Introducción

En este documento se presentan los trabajos relacionados con la analítica de datos ejecutada sobre los dos primeros pilotos, en los que se han aplicado diferentes técnicas de Machine Learning para mejorar la capacidad de detección de deterioro cognitivo a través de DigiMoca, un agente conversacional basado en el test T-MoCA. En estos dos pilotos se han recabado datos en dos centros de día diferentes para la validación del instrumento de cribado DigiMoca.

Tanto el modelo computacional utilizado como los resultados del análisis de los datos recopilados en los dos pilotos se presentan en detalle en el artículo 'Enhancing cognitive impairment screening by means of conversational agents and machile learning' al cual se puede acceder desde la sección 'Publications' en la web del proyecto Sapiens[1]. Debe ser esta publicación la que se tome como prinicipal referencia sobre los trabajos realizados. En los dos apartados siguientes se presenta un resumen tanto del modelo computacional construido como de los resultados obtenidos tras la realización de la analítica de datos.

[1] https://sapiens.webs.uvigo.es/publications/

2. Modelo computacional

El modelo computacional en el que se basa la integración de Machine Learning (ML) en DigiMoCA sigue un flujo de procesamiento de datos compuesto por cinco etapas principales. A continuación, se describe cada una de ellas:

1. Recopilación de Datos

Se obtienen datos de dos fuentes principales:

- Participantes: Resultados de pruebas cognitivas estándar como T-MoCA, MFE (Memory Failures of Everyday), y IADL (Instrumental Activities of Daily Living).
- DigiMoCA: Datos extraídos de la interacción con el agente conversacional, incluyendo:
  - Respuestas a las preguntas del test.
  - Datos adicionales como tiempos de respuesta y errores cometidos.

Estos datos son almacenados en la base de datos NoSQL de Amazon DynamoDB y descargados en formato JSON.

2. Preprocesamiento de Datos

Se realiza una limpieza y normalización de los datos para optimizar el rendimiento del modelo:

- Eliminación de características irrelevantes o redundantes (con correlación > 0.95 o varianza < 0.01).
- Normalización de datos (media 0, desviación estándar 1).
- Fusión de datos de diferentes fuentes en un solo conjunto de datos (CSV).

Se evita la fuga de datos asegurando que la normalización solo se aplica a los datos de entrenamiento y se replica en el conjunto de prueba.

3. Selección de Características (Feature Selection)

Se identifica el subconjunto óptimo de características que mejoran la precisión del modelo. Se emplea el método de importancia por permutación, que mide la caída en rendimiento al eliminar una variable:

- La variable con mayor importancia fue el número de palabras recordadas sin pistas.
- También se destacan la orientación temporal y espacial, el número de errores en tareas de atención y la cantidad de palabras generadas en una prueba de fluidez verbal.

4. Validación Cruzada y Optimización de Hiperparámetros

Se emplea la estrategia Leave-One-Group-Out (LOGO) cross-validation, en la que:

- Se separa un conjunto de prueba con un solo participante.
- Se entrena el modelo con el resto de los datos.
- Se repite el proceso hasta evaluar a todos los participantes.

Se usa optimización bayesiana para ajustar los hiperparámetros de cada modelo y mejorar su desempeño.

3. Resultados de la analítica de datos

Los resultados obtenidos en el estudio sobre la integración de técnicas de Machine Learning en el sistema DigiMoCA demuestran mejoras significativas en la detección del deterioro cognitivo leve (MCI) y en la predicción de puntuaciones en el test T-MoCA. A continuación se presentan los hallazgos más relevantes, tanto en términos cuantitativos como cualitativos.

En la tarea de predicción de la puntuación T-MoCA, se evaluaron 23 modelos de regresión, destacando los siguientes por su rendimiento superior:

- El modelo de Regresión Lineal alcanzó un coeficiente de determinación (R^2) de 0.78 con un error cuadrático medio (RMSE) de 1.90.
- El modelo Ridge Regression mostró una ligera mejora, con un R^2 de 0.79 y un RMSE de 1.88.
- El modelo Random Forest obtuvo un R^2 de 0.77 y un RMSE de 1.92.
- El mejor resultado fue obtenido con el modelo Voting Regressor, que combina las fortalezas de otros modelos lineales, alcanzando un R^2 de 0.80 y reduciendo el error a un RMSE de 1.85.

En la tarea de clasificación del deterioro cognitivo leve (MCI), los resultados también fueron notables. El modelo de Regresión Logística alcanzó un área bajo la curva ROC (AUC) de 0.92, con una sensibilidad del 91% y una especificidad del 83%. El modelo Support Vector Machine (SVM) logró un AUC de 0.93, acompañado de una sensibilidad del 93% y una especificidad del 84%. Por otro lado, el modelo Multi-Layer Perceptron (MLP) obtuvo un AUC de 0.92, con una sensibilidad del 92% y una especificidad del 85%.

El mejor rendimiento en la detección de MCI se obtuvo con el modelo Voting Classifier, que combina la fuerza de los modelos mencionados anteriormente. Este algoritmo alcanzó un AUC de 0.95, junto con una sensibilidad del 94% y una especificidad del 85%, superando así significativamente a los métodos tradicionales como el Mini-Mental State Examination (MMSE) y el propio DigiMoCA sin machine learning.

El análisis cualitativo reveló que la inclusión de ciertas características extraídas durante la administración de DigiMoCA fue clave para el éxito del modelo. En particular, se observó que el número de palabras recordadas sin pistas fue el indicador más relevante para predecir la puntuación T-MoCA, seguido de la orientación temporal y espacial y los errores cometidos en tareas de atención. Estas variables, que no suelen ser consideradas en las pruebas cognitivas tradicionales, resultaron determinantes para mejorar la precisión de los modelos.

Otro hallazgo importante fue que el tiempo de respuesta de los participantes, una variable adicional capturada por DigiMoCA, contribuyó significativamente a incrementar la precisión del sistema en la clasificación del deterioro cognitivo. Este resultado destaca el potencial de DigiMoCA para detectar signos tempranos de deterioro cognitivo leve que podrían pasar desapercibidos en evaluaciones tradicionales.

En conjunto, los resultados obtenidos demuestran que la integración de técnicas de Machine Learning en DigiMoCA permite obtener resultados más precisos en la predicción del deterioro cognitivo leve y en la estimación de puntuaciones en el test T-MoCA. El enfoque basado en aprendizaje automático no solo mejora el rendimiento del sistema, sino que también automatiza la evaluación, lo que facilita su uso en entornos no clínicos y amplía el acceso a herramientas de detección temprana de deterioro cognitivo.

4. Personal participante

Equipo de investigación:

- Investigadores Principales (IPs)
  - Luis Eulogio Anido Rifón (Uvigo)
  - Luis Modesto Álvarez Sabucedo (Uvigo)
- Equipo de Investigación (EI)
  - Manuel Fernández Iglesias (Uvigo)
  - Martín Llamas Nistal (Uvigo)
  - Manuel Gandoy Crego (USC)
  - Marcos Pazos Couselo (USC)
  - Manuel Caeiro Rodríguez (Uvigo)
  - Juan Manuel Santos Gago (Uvigo)
  - Fernando Ariel Mikic Fonte (Uvigo)
- Equipo de Trabajo (ET)
  - Alba Elena Martínez Santos (USC)
  - César Bugallo Carrera (USC)
  - Sonia María Valladares Rodríguez (USC)
  - Martín Liz Domínguez (Uvigo)
  - Moises Ruben Pacheco Lorenzo (Uvigo)
- Personal Contratado (PC)
  - Iván Otero González
  - Noelia Lago Priego