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.4.3: 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. Entrada de Datos 3
2. Preprocesamiento de Datos 3
3. Modelos de Clasificación 4
4. Evaluación y Optimización 4
5. Implementación en la Nube 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 el tercer piloto, en los que se han aplicado diferentes técnicas de Machine Learning para mejorar la capacidad de detección de depresión en personas mayores a través de DigiYesavage, un agente conversacional basado en el test GDS-15. En este tercer piloto se han recabado datos en tres centros de día diferentes para la validación del instrumento de cribado DigiYesavage

Tanto el modelo computacional utilizado como los resultados del análisis de los datos recopilados en los dos pilotos se presentan en detalle en el capítulo 6 de la tesis 'Agentes conversacionales inteligentes para la detección de depresión' al cual se puede acceder desde la sección 'Publications' en la web del proyecto Sapiens[1]. Debe ser ese documento la que se tome como principal 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 DigiYesavage está compuesto por varias capas y módulos interconectados, que permiten capturar datos, procesarlos y clasificarlos utilizando técnicas de aprendizaje automático.

1. Entrada de Datos

El modelo recibe información a través de la interacción por voz con la skill DYes en Amazon Alexa.
Los datos de entrada se dividen en tres categorías:

- Respuestas al cuestionario GDS-15 (Yesavage)
  - Preguntas con respuesta binaria (*Sí/No*), utilizadas para calcular la puntuación de depresión.
- Datos de interacción
  - Tiempo de respuesta de cada usuario (en segundos).
  - Número de repeticiones necesarias para responder una pregunta.
  - Interrupciones o silencios prolongados.
- Datos sociodemográficos
  - Edad, género, nivel educativo.
  - Estado civil y situación de convivencia (viviendo solo o acompañado).

Estos datos se almacenan en una base de datos DynamoDB de Amazon y se transmiten al backend para su procesamiento.

2. Preprocesamiento de Datos

Una vez capturados, los datos pasan por un proceso de limpieza y transformación para su uso en los modelos de clasificación:

1. Codificación de variables categóricas
   - Respuestas binarias (Sí/No) transformadas en valores 0 y 1.
   - Variables como género y estado civil convertidas en representaciones numéricas mediante *one-hot encoding*.
2. Normalización y escalado
   - Se estandarizan variables continuas (edad, tiempos de respuesta) para evitar sesgo en los modelos.
3. Manejo de valores faltantes

   - Se imputan valores ausentes mediante la media o la moda, dependiendo del tipo de variable.
4. División del dataset
   - Se separa en conjunto de entrenamiento (80%) y prueba (20%), manteniendo la proporción de clases mediante estratificación.

3. Modelos de Clasificación

El modelo DYes implementa ocho algoritmos de clasificación, evaluados para determinar cuál ofrece el mejor rendimiento.
Estos modelos incluyen:

- Regresión Logística (modelo principal):
  - Proporciona interpretabilidad y precisión en la clasificación.
- Random Forest y AdaBoost:
  - Mejoran la robustez del modelo mediante agregación de árboles de decisión.
- Support Vector Machines (SVM):
  - Destacado en discriminación de clases con un área bajo la curva ROC alta.
- Red Neuronal (MLP), Gradient Boosting y KNN:
  - Modelos explorados, pero con menor rendimiento en esta tarea.

Cada modelo es entrenado y evaluado con métricas como precisión, recall, F1-score y curva ROC.

4. Evaluación y Optimización

- Se utiliza Validación Cruzada (k-fold 10) para mejorar la estabilidad del modelo.
- Se aplica Permutación de Importancia de Características para identificar las variables con mayor peso en la clasificación.
- Los modelos se comparan y se selecciona el que maximiza la precisión y minimiza los falsos negativos.

5. Implementación en la Nube

- Backend en AWS Lambda (Python) para ejecutar la lógica de clasificación.
- DynamoDB como base de datos escalable para almacenar interacciones y resultados.
- Skill de Alexa integrada con el modelo para ofrecer una interfaz accesible y natural.

3. Resultados de la analítica de datos

El análisis de datos del clasificador DYes se centró en evaluar la efectividad de ocho modelos de aprendizaje automático en la detección de síntomas depresivos en adultos mayores. Se midieron diferentes métricas de clasificación para determinar cuál de los modelos lograba una mayor precisión y un mejor desempeño en la discriminación de casos positivos de depresión. Entre los modelos evaluados, la Regresión Logística demostró ser el más eficiente, con una precisión del 84% y un F1-score de 0.77, lo que indica un adecuado balance entre la capacidad de predecir correctamente los casos positivos y minimizar los errores de clasificación. Otros modelos como Random Forest y AdaBoost lograron un rendimiento cercano, con una precisión del 79% y valores similares en recall y precisión. Por otro lado, modelos más complejos como Support Vector Machines (SVM) destacaron en la discriminación de clases, con un área bajo la curva ROC (AUC) de 0.89, lo que sugiere una alta capacidad para diferenciar entre individuos con y sin síntomas depresivos, aunque con una precisión global ligeramente menor. Modelos basados en aprendizaje profundo, como las Redes Neuronales Multicapa (MLP) y K-Nearest Neighbors (KNN), presentaron un rendimiento inferior, con una precisión del 74%, debido en parte a la cantidad limitada de datos utilizados para el entrenamiento.

El análisis de la matriz de confusión confirmó que la Regresión Logística tuvo el menor número de errores, con solo tres clasificaciones incorrectas en la muestra. Esto significa que es el modelo más confiable para identificar la presencia de síntomas depresivos en adultos mayores. Sin embargo, se observó que modelos como Random Forest y AdaBoost, aunque con menor precisión general, mostraron un mejor desempeño en términos de robustez frente a posibles variaciones en los datos de entrada. Además, los modelos de boosting, como Gradient Boosting, lograron una buena capacidad predictiva, pero con un mayor costo computacional, lo que los hace menos prácticos para una implementación en entornos de evaluación rápida.

Además de la precisión de los modelos, el estudio también evaluó la relación entre la depresión y la presencia de deterioro cognitivo leve (DCL), lo que permitió clasificar a los participantes en diferentes grupos. Se encontró que aproximadamente el 41% de los participantes presentaban signos de demencia, con una edad promedio de 79 años, mientras que el 59% restante no mostraban signos de deterioro cognitivo significativo y tenían una edad media de 72 años. Estos resultados sugieren que la edad es un factor determinante en la aparición de síntomas depresivos y su posible relación con trastornos neurodegenerativos. Asimismo, el análisis mostró que los tiempos de respuesta y el número de repeticiones en la interacción con el agente conversacional DYes eran significativamente mayores en los participantes con demencia, lo que respalda la idea de que la velocidad de procesamiento cognitivo puede ser un indicador clave en la detección temprana de deterioro mental.

Dentro del análisis de importancia de variables, se identificaron cinco factores clave que influyen en la clasificación de los pacientes con síntomas depresivos. En primer lugar, la puntuación en la Escala de Depresión Geriátrica (GDS-15) fue la variable con mayor peso en la predicción, lo que confirma la validez de este instrumento para evaluar síntomas depresivos en adultos mayores. En segundo lugar, se encontró que el tiempo de respuesta a las preguntas del test es un factor crítico, ya que los pacientes con depresión o deterioro cognitivo tienden a tardar más en responder. También se destacó el número de repeticiones necesarias para completar el cuestionario, ya que los individuos con dificultades cognitivas requerían que las preguntas fueran formuladas varias veces antes de responder. Además, variables demográficas como la edad, el estado civil y el nivel educativo influyeron en la clasificación, lo que sugiere que aspectos sociales y culturales pueden desempeñar un papel en la detección de la depresión en este grupo poblacional.

En conclusión, los resultados del análisis de datos respaldan la efectividad del clasificador DYes para la detección de síntomas depresivos en adultos mayores, con la Regresión Logística como el modelo más eficiente para esta tarea. Se confirma que el uso de técnicas de aprendizaje automático, combinado con un agente conversacional basado en voz, puede proporcionar una herramienta precisa y accesible para la evaluación de la salud mental.

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