Generación de un modelo de predicción y categorización

Andrés Alejandro Guzmán González[A01633819], Ernesto Reynoso Lizárraga[A01639915], Joel Isaias Solano Ocampo[A01639289], Luis Rodolfo Bojorquez Pineda[A01250513], and Tania Sayuri Guizado Hernandez[A01640092]

Tecnológico de Monterrey, Campus Guadalajara, Av. General Ramón Corona 2514, Nuevo México, 45201 Zapopan, Jalisco, México

Resumen Este informe presenta nuestra estrategia para desarrollar un modelo de inteligencia artificial destinado a predecir categorías a partir de datos de aceleración en tres ejes de dos acelerómetros ubicados en la espalda baja y el muslo derecho de adultos mayores por un timepo aproximado de 40 minutos. Utilizamos un conjunto de datos compuesto por 15 muestras obtenidas del conjunto Activity Recognition in Senior Citizens [3]. Detallamos los resultados de nuestros análisis exploratorios y cómo estos condujeron a nuestra solución final: un modelo de clasificación basado en Random Forest. Nuestro enfoque se centra en la predicción de las actividades identificadas en el conjunto de datos para el monitoreo de movimientos de adultos mayores, lo que tiene aplicaciones significativas en el campo de la atención médica y la calidad de vida de esta población en crecimiento. Este trabajo contribuye a la búsqueda de estrategias para la asistencia médica de adultos mayores.

Keywords: Data set · Modelo · Clasificador · Ajuste · Predicción · Categorización · Random Forest · Regresión Logística.

1. Análisis Exploratorio de Datos

El primer proceso realizado fue un análisis exploratorio de los datos, esto con la finalidad de ver cómo se comportan los datos y poder identificar si existen datos faltantes, datos atípicos, etc. Por ello nos dimos a la tarea de revisar cómo se presentaban los datos y lo que representaban en primer instancia.

Cuadro 1: Ejemplos de los valores del conjunto de datos. [3]

timestamp	back_x	back_y	back_z	thigh_x	thigh_y	thigh_z	label
43:32.6	-0.989502	-0.045166	-0.193115	-0.984131	-0.144043	0.000977	1
43:02.1	-0.899658	0.081055	0.080322	-0.968994	0.030762	-0.102783	3
16:57.3	-0.817871	0.035156	-0.259277	-0.946289	-0.180664	-0.236572	4
53:31.4	-1.005371	0.09375	-0.295166	-1.080811	0.012451	-0.112061	5
43:40.7	-0.986816	-0.187012	0.135254	0.06543	0.096191	-1.257324	7

Los datos de la muesta representan:

- **timestamp**: tiempo en el que se tomó la muestra.
- **back_x**: aceleración en el eje X del acelerómetro ubicado en la espalda.
- **back_y**: aceleración en el eje Y del acelerómetro ubicado en la espalda.
- **back_z**: aceleración en el eje Z del acelerómetro ubicado en la espalda.
- **thigh_x**: aceleración en el eje X del acelerómetro ubicado en el muslo.
- **thigh_y**: aceleración en el eje Y del acelerómetro ubicado en el muslo.
- **thigh_z**: aceleración en el eje Z del acelerómetro ubicado en el muslo.
- **label**: actividad realizada por el adulto mayor.

Y a su vez los valores de las actividades en la columna label representan:

- **1**: Caminar (walking).
- **3**: Arrastramiento (shuffling).
- **4**: Subir escaleras (stairs (ascending)).
- **5**: Bajar escaleras (stairs (descending)).
- **6**: Mantenerse de pie (standing).
- **7**: Sentarse (sitting).
- **8**: Acostarse (lying).

Con este contexto se procedió a realizar gráficas de cada una de las variables de aceleración con respecto de los índices de tiempo, esto con la finalidad de ver de forma visual el comportamiento de la aceleración, tal como se muestra en la figura 1.

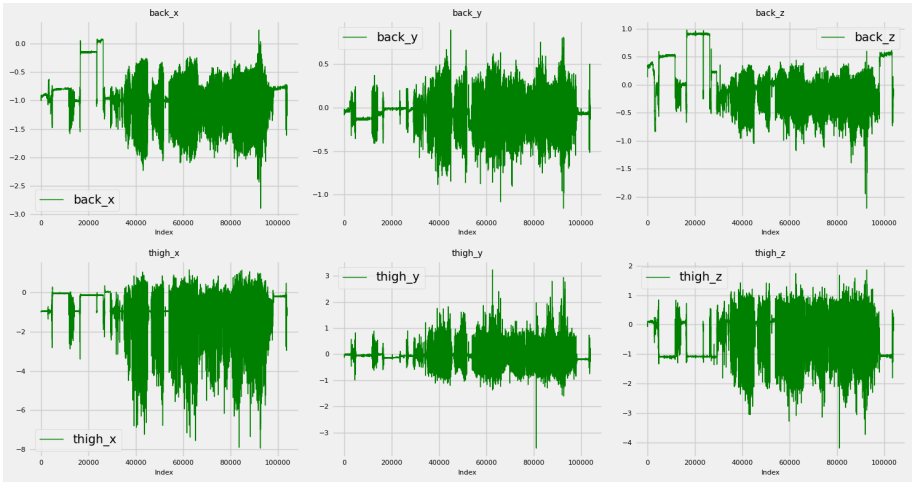


Figura 1: Gráfias aceleración vs índice de tiempo.

Considerando el volumen de los datos no es fácil de interpretar la información mostrada; sin embargo, al considerar que las muestra ya venían categorizadas,

tomamos la decisión de marcar las fronteras de cada una de las actividades, en una gráfica para visualizar de mejor manera el comportamiento de la aceleración en cada una de las actividades, tal como se muestra en la figura 2.

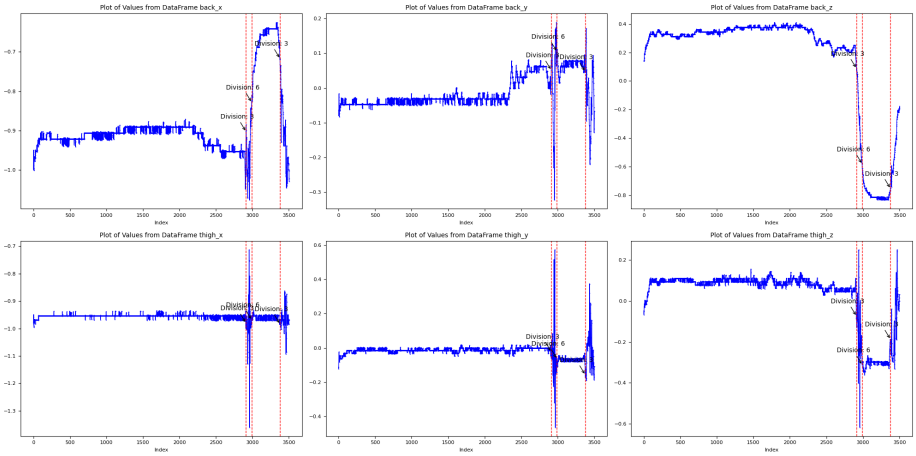


Figura 2: Gráfias aceleración vs índice de tiempo.

Con esta nueva visualización se llegó a la conclusión de que la aceleración de acuerdo con la actividad que estuviera realizado el adulto mayor; se mantenía en un rango un tanto similar. Sin embargo, cabe mencionar que para este momento la estrategia de solución ya consideraba; la necesidad de generar un modelo de clasificación puesto que la variable de respuesta es categórica. También se identificó que los conjuntos de datos no siempre contenían muestras para todas las categorías de actividades; destacando la necesidad de tener el combinar las muestras para evitar que el modelo no tuviera información suficiente para categorizar los datos entrantes. Finalmente, se determinó que la variable tiempo no era relevante para la solución del problema, puesto que la base del modelo se centraría en el análisis de los valores de la aceleración en cada uno de los ejes para hacer la categorización correspondiente.

Con estas conclusiones, se procedió a realizar un análisis de correlación entre las variables para poder detonar las estrategias que nos permitirán construir el modelo matemático para eventualmente entrenar a la inteligencia artificial. Cabe mencionar que esta prueba nos permite identificar la necesidad de realizar un proceso de estandarización de los datos.

El resultado del análisis de correlación se mostró en una matriz de correlación de variables; en la pudo observar que para algunas de estas, hubo valores de correlación significativamente altos, destacando la necesidad de escalar los datos al momento de empezar a generar el modelo. La matriz de correlación resultante se muestra en la figura 3.

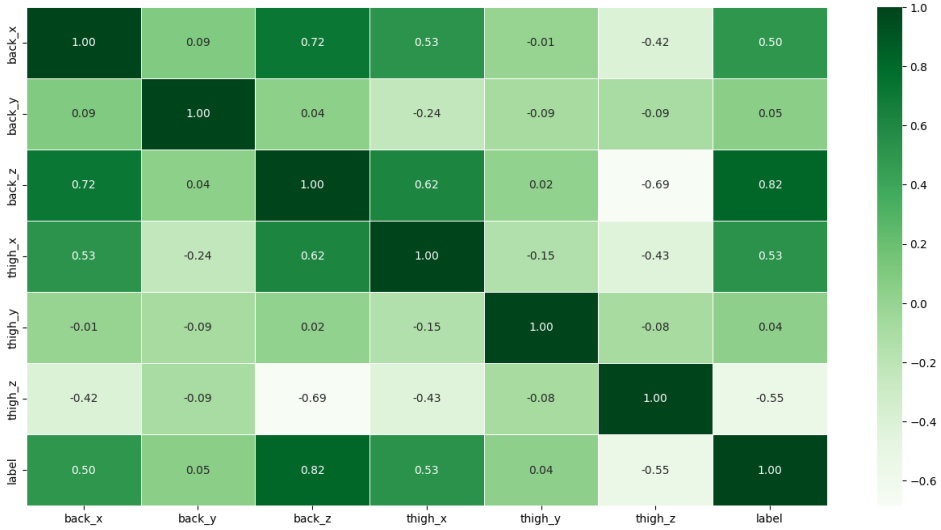


Figura 3: Matriz de correlación entre variables.

2. Metodología

En primer lugar, se consideró la idea de tener una referencia espacial de los valores de la aceleración en cada uno de los ejes, por lo que se optó por buscar la doble integral de las aceleraciones en cada uno de los ejes, esto con la finalidad de obtener la posición del adulto mayor con el paso del tiempo. Sin embargo, al realizar la doble integral y analizar el resultado no hizo sentido para el equipo y sobre todo para la solución de modelo el tener la referencia espacial, puesto que la solución se centraría en la clasificación de las actividades que realizaba el adulto mayor. Por lo que se optó por volver a la propuesta inicial de generar un modelo de clasificación considerando solo los rangos en los que se encontraba la aceleración en los ejes X, Y y Z con el paso del tiempo.

Con ello se procedió a seleccionar 4 modelos de clasificación *Regresión Logística, Random Forest, SVC - Linear y Regresión Lineal.* Se optó por estos modelos, puesto que son los que se han visto en clase y se consideró que serían los más adecuados para una fase de experimentación.

El proceso que se siguió para diseñar los modelos se destaca en los siguientes puntos:

- Se concatenaron los datos de las 15 muestras.
- Se realizó un proceso de selección de variables.
- Se estandarizaron los datos.
- Se realizó evaluó la necesidad de balancear las clases.
- Se realizó un proceso de entrenamiento de los modelos.
- Se realizó un proceso de validación de los modelos.
- Se realizó la selección del mejor modelo.

3. Experimentos

Tal como se mencionó anteriormente, se concatenaron las muestras del data set; esto con la necesidad de tener muestras en todas las posibles acciones o categorías resultantes. Respecto a la selección de variables. La variable de respuesta corresponde a 'label' (que es una variable categórica) y se consideraron variables del modelo a las aceleraciones en los ejes X, Y y Z de los dos acelerómetros ubicados en la espalda y el muslo.

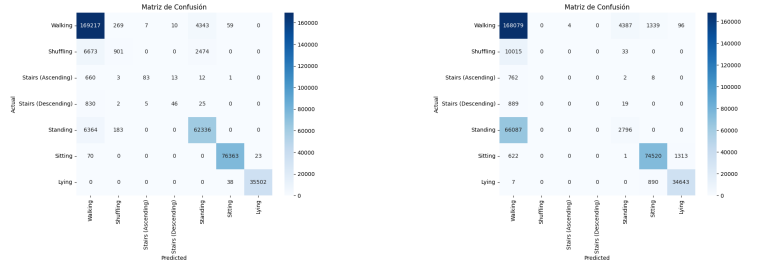
Para el proceso de estandarización de los datos se utilizó la función *StandardScaler* de la librería *sklearn.preprocessing*.

En primera instancia los modelos se corrieron con los clasificadores *Regresión Logística, Random Forest, SVC - Linear y Regresión Lineal.* y sin un balanceo de datos. Cabe mencionar que para el caso de *SVC - Linear y Regresión Lineal.* no se obtuvieron resultados pues el tiempo de ejecución fue muy largo y eventualmente marcaron errores de memoria. Por ello se optó por eliminarlos de la lista de modelos a considerar. En caso contrario, los modelos *Regresión Logística y Random Forest* se ejecutaron correctamente y se obtuvieron resultados favorables, los cuáles se muestran en la siguinte tabla.

Cuadro 2: Resultados de clasificación

Clasificador	Exactitud	Prom-Macro	Prom-Ponderado
Regresión Logística	0.76	0.54	0.69
Random Forest	0.94	0.61	0.93

Si bien el valor de la exactitud es 0.9398 en el caso del clasificador *Random Forest*, lo que significa que nuestro modelo ha clasificado correctamente aproximadamente el 94 % de las muestras en el conjunto de prueba. Es un indicador positivo, pero la exactitud puede ser engañosa o errónea si las clases están desequilibradas.



(a) Random Forest. (b) Regresión Logística.

Figura 4: Matrices de confusión de los modelos.

Las matrices de confusión de los modelos muestran la distribución de las predicciones del modelo en cada clase real. Tal como se muestra en la figura 4; el modelo *Random Forest* tiene un mejor desempeño que el modelo *Regresión Logística*.

- Para la clase 1 *walking* el modelo predijo correctamente 169,217, pero hay un total de 46,88 falsos negativos (clasificaciones en otras clases que realmente pertenecían a la clase 1).
- Para la clase 3 *shuffling* el modelo predijo correctamente 901, pero hay un total de 9,147 falsos negativos.
- Para la clase 4 *stairs (ascending)* el modelo predijo correctamente 83, pero hay un total de 689 falsos negativos.
- Para la clase 5 *stairs (descending)* el modelo predijo correctamente 46, pero hay un total de 862 falsos negativos.
- Para la clase 6 *standing* el modelo predijo correctamente 62,336, pero hay un total de 6,547 falsos negativos.
- Para la clase 7 *sitting* el modelo predijo correctamente 76,363, pero hay un total de 93 falsos negativos.
- Para la clase 8 *lying* el modelo predijo correctamente 35,502, pero hay un total de 38 falsos negativos.

En general, este modelo parece ser muy preciso para algunas clases, pero tiene dificultades para clasificar correctamente otras. Por lo que se puede considerar abordar el desequilibrio de clases o incluso ajustar los hiperparámetros del modelo. Sin embargo, se optó por realizar un proceso de balanceo de clases analizando en primer lugar la distribución de las clases en el conjunto de datos, tal como se muestra en la figura 5.

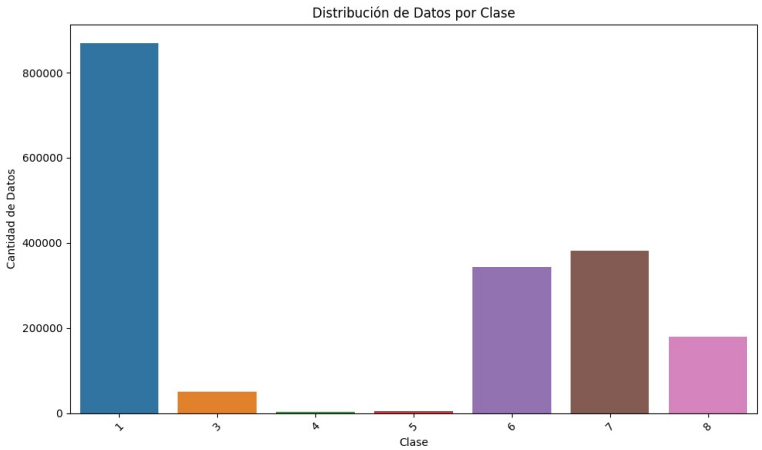
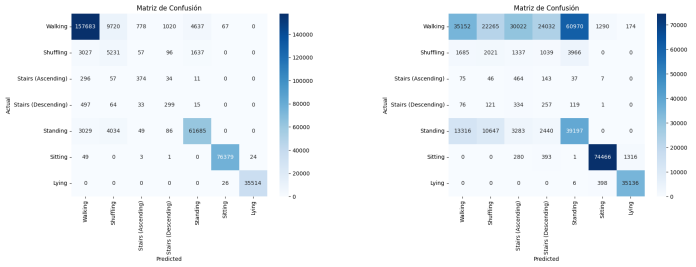


Figura 5: Balance de clases.

Para el proceso de balanceo de clases se utilizó la función *SMOTE* de la librería *imblearn.over_sampling*. Una vez hecho el balanceo de clases se procedió a realizar el proceso de entrenamiento y validación de los modelos, nuevamente los cuales arrojaron los siguientes resultados.

Cuadro 3: Resultados de clasificación

Clasificador	Exactitud	Prom-Macro	Prom-Ponderado
Regresión Logística	0.51	0.41	0.53
Random Forest	0.92	0.69	0.93



(a) Random Forest. (b) Regresión Logística.

Figura 6: Matrices de confusión de los modelos.

Estos resultados muestran nuevamente que el modelo de clasificación con *Random Forest* tiene un mejor desempeño que el modelo de *Regresión Logística*, el cual empeoró al hacer el balanceo de clases. Por ello se optó por optimizar la solución utilizando el clasificador *Random Forest* pues entre sus características se encuentra la capacidad de manejar múltiples características, manejo de relaciones no lineales y datos desbalanceados también su naturaleza lo hace menos propenso al sobre ajuste pues combina múltiples árboles[1].

Además, en el caso de la regresión logística, la clasificación de las clases 3,4,5 y 6 se concentraron en la clase 1, lo que significa que el modelo no es capaz de distinguir entre estas clases. Por lo que se optó por descartar este modelo.

4. Resultados

Considerando los hallazgos de los experimentos se decidió continuar con el modelo de clasificación *Random Forest* se procedió a buscar estrategias para mejorar el desempeño del modelo. Si bien un buen indicador de desempeño es la exactitud, se optó por priorizar el número de verdaderos positivos en las predicciones, puesto que el objetivo es clasificar correctamente las actividades.

Retomando los resultados de las matrices de confusión, se observó que el modelo tiene dificultades para clasificar actividades que en cuestión numérica son muy similares, tal es el caso de la clase 1 *walking* y la clase 3 *shuffling* así como las clases 4 y 5 *stairs (ascending) y stairs (descending)*. Identificar esta relación llevó a un proceso de agrupación de clases, por lo que la clase 1 fue renombrada a moving y la clase 4 a stairs tal como se muestra en la figura 7.

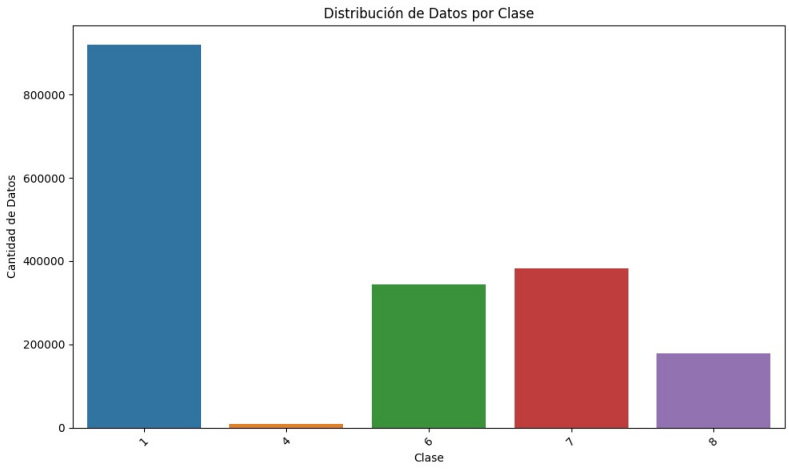


Figura 7: Balance de clases después de agrupar.

Con esta estrategia, se procedió a realizar nuevamente el proceso de balanceo de clases y entrenamiento del modelo de clasificación, el cual obtuvo una exactitud del 98 % y al validar la distribución de verdaderos positivos el resultado es extremadamente mejor tal como se muestra en la figura 8.

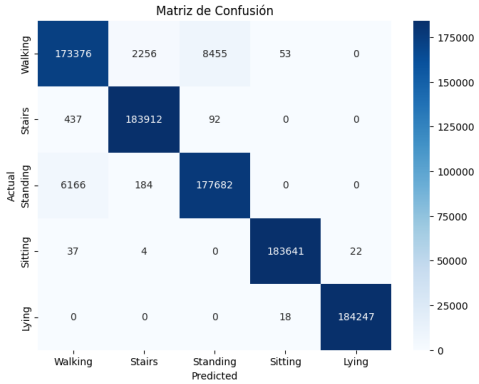


Figura 8: Balance de clases después de agrupar.

Tal como se presentó anteriormente, se puede decir que el modelo es muy bueno considerando que la mayoría de los datos se encuentran en la diagonal de la matriz de confusión y que los falsos positivos y falsos negativos son muy pocos en comparación con el número de verdaderos positivos. Por lo que se considera que el modelo es capaz de clasificar correctamente, las actividades que realiza el adulto mayor después de realizar la agrupación de clases.

5. Conclusiones

Con estos resultados se logró cumplir con el objetivo de generar un modelo de clasificación que permita clasificar las actividades realizadas por un adulto mayor. Sin embargo, consideramos que el modelo puede ser mejorado. Por ejemplo a consecuencia de la agrupación de clases, se perdió la capacidad de clasificar la actividad de arrastrarse y subir o bajar escaleras.

Con estas nuevas limitantes una opción sería el aumentar el tamaño de las muestras recopiladas en el data set. Si bien este tenía un total 15 muestras con un gran conjunto de datos tomados en tiempo real. La gran diferencia en la distribución de las actividades fue un gran reto, pues a pesar del balanceo los modelos no lograron aumentar significativamente su exactitud.

Como equipo se llegó a la conclusión que una forma de optimizar el modelo presentado y asumiendo que se pueda aumentar el tamaño del data set, sería el diseñar un modelo jerárquico de clasificación, en el cual se tenga un modelo que clasifique las actividades de forma general tal y como se presentó en la sección de resultados y posteriormente se tengan modelos que clasifiquen las actividades de forma más específica. Por ejemplo, el modelo final es capaz de identificar si el adulto mayor está en las escaleras, un segundo modelo haría la clasificación entre el asenso y descenso. Lo mismo pasaría con el arrastramiento y el caminar.

6. Conclusiones Individuales

Andrés Alejandro Guzmán González

Con el desarrollo del presente trabajo tuve la oportunidad de poner a prueba los conocimientos adquiridos en clase, lo cuales van desde el realizar un análisis exploratorio de los datos, validad la necesidad de hacer un balanceo, priorizar tipo de clasificadores así como la implementación, entrenamiento y validación de un modelo de clasificación. Considero que el mayor reto fue el poder determinar la estrategia a seguir para la solución para la solución del problema, pues como se mencionó anteriormente, se consideró la posibilidad de generar un modelo que pudiera también darnos una idea de la posición del adulto mayor con el paso del tiempo. Sin embargo, no se obtuvo un resultado que tuviera sentido.

Por otro lado, considero que el mayor aprendizaje fue el poder identificar la posibilidad de realizar un proceso de agrupación de clases, pues al realizar el análisis exploratorio un poco más profundo de los datos se identificó que las actividades de caminar y arrastrarse tenían parámetros similares; lo mismo

pasaba con las actividades de subir y bajar escaleras. Si bien, este proceso generó un grado de incertidumbre el poder validarlo con el modelo de clasificación y ver que el resultado fue favorable, me permitió a mí y al equipo el poder tener una idea de cómo se puede mejorar el modelo en un futuro, así como a identificar nuevas estrategias para hacer un modelo más óptimo y acertado.

Ernesto Reynoso Lizárraga

A lo largo del desarrollo de este reto y primera parte de la concentración de Inteligencia Artificial considero que una de las actividades mas importantes que se realizaron fue la exploración y entendimiento de los datos con los que estábamos trabajando, esto nos facilita mucho saber qué tipo de métodos, técnicas y procedimientos tenemos que emplear para llegar al objetivo del reto con la mejor efectividad posible. Además, gracias a esta exploración pudimos identificar comportamientos en ciertas actividades que, consultando con los profesores, pudimos decidir la mejor estrategia para mejorar la precisión del modelo.

Otra actividad que implementamos en este reto fue la implementación de otros modelos con distintos métodos de clasificación, esto para buscar otro modelo que pudiese ajustarse mejor a los datos. Sin embargo, ninguno se ajusto de manera tan precisa como el modelo de Random Forest, de hecho, dos de los modelos adicionales que probamos nos daban un resultado extremadamente deficiente y debido al tiempo con el que contamos no pudimos explorar más en los modelos, por lo que estos quedaron como meras observaciones.

Joel Isaias Solano Ocampo

Para nuestro modelo de AI utilizado para la solución de nuestro reto en el bloque 1 tuvimos que analizar una gran cantidad de datos que hacen referencia a las diferentes aceleraciones medidas en unidades de gravedad captadas por 2 sensores que fueron colocados en la espalda y la pierna de personas mayores para comprender mejor su comportamiento al momento de moverse. Cabe recalcar que nunca antes yo había trabajado con bastantes datos y podemos decir que fue la 1er cosa a la que me gustaría hacer hincapié con mi conclusión: realmente nuestra aproximación inicial fue de forma individual, es decir, que evaluamos los datos de persona a persona y no todos los datos al mismo tiempo; esto ayudo a no trabajar con tantos datos y a la vez, comprender mejor la naturalidad de los mismos, tener un mejor contexto de su propio comportamiento de las muestras y encontrar patrones para eventualmente, juntar todos los datos y evaluarlos en conjunto.

Conforme fueron avanzando las clases, el modelo de AI para la solución de nuestro reto empezó a cobrar mas sentido ya que aprendíamos conceptos, técnicas, modelos y entre otras estrategias para lograr la predicción de movimiento en personas mayores con los muestreos que teníamos a nuestra disposición. Fue interesante replantearnos constantemente la implementación del modelo en diferentes facetas ya que había situaciones donde cuando considerábamos a primera instancia que una solución era la más optima, podía llegar una clase que nos

podría sugerir lo que realmente estaba pasando con nuestro modelo en general y que la supuesta solución que teníamos en mente no era la aproximación más adecuada. A final de cuentas, me encuentro muy satisfecho con la implementación final que realizo mi equipo y yo en conjunto, realizamos un modelo muy diferente a nuestros demás compañeros y me gusto ver que los demás equipos dieron con otros panoramas viables o simplemente diferentes a lo nuestro.

Luis Rodolfo Bojorquez Pineda

En este informe, presentamos nuestra estrategia para desarrollar un modelo de inteligencia artificial destinado a predecir categorías a partir de datos de aceleración en tres ejes de dos acelerómetros ubicados en la espalda baja y el muslo derecho de adultos mayores durante un período de tiempo aproximado de 40 minutos. A través de un análisis exploratorio de datos, se identificaron patrones en las aceleraciones que llevaron a la selección de un modelo de clasificación basado en Random Forest. Abordamos el desafío del desequilibrio de clases mediante técnicas de balanceo y consideramos la agrupación de clases similares para mejorar el rendimiento del modelo. Nuestro enfoque tiene aplicaciones significativas en la atención médica y la calidad de vida de los adultos mayores, contribuyendo a la búsqueda de estrategias para su asistencia. A pesar de los desafíos encontrados, hemos logrado un modelo prometedor que puede mejorarse aún más con la expansión del conjunto de datos y la implementación de un modelo jerárquico de clasificación. Este trabajo sienta las bases para futuros avances en el monitoreo de la actividad de los adultos mayores.

Tania Sayuri Guizado Hernandez

En el desarrollo de la implementación de nuestro modelo de inteligencia artificial, a partir de la base de datos Activity Recognition in Senior Citizens, nos aseguramos de llevar a la práctica conocimientos relevantes adquiridos para un correcto funcionamiento en la clasificación de las actividades categorizadas del conjunto de datos. En nuestro caso, el modelo de clasificación seleccionado fue Random Forest y su uso respectivamente, garantizo que no hubiera sesgos indeseados y que las decisiones tomadas por el modelo sean justas y precisas. Además, abordar el desequilibrio de clases mediante técnicas de balanceo y considerado la agrupación de clases similares mejoró la fiabilidad del modelo. Personalmente, valoro enormemente este proceso, ya que se demostró un compromiso sólido con la búsqueda de la máxima precisión en nuestras predicciones mediante decisiones bien fundamentadas.

Por otro lado, este proyecto tiene una significancia importante porque ejemplifica como la tecnología puede aplicarse de manera ética y beneficiosa para predecir aspectos relevantes, en este caso en el monitoreo de la actividad de adultos mayores, lo que podría atribuir a su calidad de vida y atención medica. En modo de conclusión, es valioso reconocer que aunque no siempre sea posible

de lograr el continuo esfuerzo por mejorar un modelo se traducirá en predicciones más precisas y, en última instancia, en un mayor aporte a los resultados prácticos.

Referencias

1. Geron, A.: Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly (2019)
2. Documentación de Imbalanced-learn. Disponible en: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html. Accedido en agosto de 2023.
3. Logacjov, A. Ustad,A. *Kaggle*. Disponible en: https://www.kaggle.com/datasets/anshtanwar/adult-subjects-70-95-years-activity-recognition. Accedido el 31 de agosto, 2023.