

Anexos

Anexo A: Resultados complementarios

A.1 Resultados utilizando los datos de la visita en baseline

El comportamiento más destacable es el que se da en la métrica sensitivity, debido a los resultados del modelo entrenado con EBM_set, ya que en media es prácticamente el que peores resultados obtiene. El modelo entrenado con RSF_set es el que mejor resultado presenta, el segundo modelo con mejores resultados es el entrenado con ADNIMERGE set.

Se puede concluir que el modelo más robusto es el entrenado con RSF_set ya que en cuatro de las cinco métricas es el que mejores resultados presenta en media, y en el caso de specificity es el segundo con mejores resultados. Asimismo, el buen rendimiento que presenta para sensitivity no implica un rendimiento inferior en specificity en comparación con el resto de los modelos, como sí ocurre para los modelos entrenados con EBM_set y ADNIMERGE_set.

Por otro lado, los modelos entrenados con EBM_set y ADNIMERGE_set presentan comportamientos similares en términos generales, a excepción de las métricas sensitivity y specificity donde presentan resultados contrarios. En el caso del modelo entrenado con ADNIMERGE_set los resultados en sensitivity son superiores, mientras que los resultados en specificity son superiores para el EBM_set. El modelo entrenado con MinimalRNN_set presenta en términos generales peores resultados que estos dos modelos.

| Años | Modelos | | | | |
|-------|------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 1 | 79.91 \pm 3.54 | 84.38 \pm 3.03 | 85.43 \pm 5.09 | 86.00 \pm 2.56 | 82.88 \pm 3.38 |
| Año 2 | 71.91 \pm 5.69 | 76.91 \pm 3.61 | 77.81 \pm 3.41 | 78.66 \pm 4.72 | 76.59 \pm 4.23 |
| Año 3 | 73.57 \pm 5.00 | 78.78 \pm 3.97 | 82.21 \pm 3.68 | 78.60 \pm 4.18 | 78.40 \pm 4.53 |
| Año 4 | 78.50 \pm 1.39 | 81.76 \pm 6.46 | 84.20 \pm 2.14 | 83.53 \pm 5.64 | 84.41 \pm 3.39 |
| Año 5 | 76.88 \pm 3.81 | 82.91 \pm 3.73 | 81.62 \pm 5.23 | 84.41 \pm 4.17 | 81.41 \pm 2.82 |
| Media | 76.16 | 80.95 | 82.25 | 82.24 | 80.74 |

Tabla A.1: Media y desviación típica de los distintos modelos para accuracy

| Años | Modelos | | | | |
|-------|------------------|-------------------------------------|-------------------|-------------------------------------|-------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 1 | 42.16 \pm 7.30 | 42.02 \pm 12.91 | 47.08 \pm 15.36 | 48.59 \pm 12.84 | 42.17 \pm 11.66 |
| Año 2 | 66.29 \pm 6.76 | 69.37 \pm 6.43 | 62.59 \pm 13.40 | 69.81 \pm 5.58 | 71.25 \pm 4.28 |
| Año 3 | 72.32 \pm 7.97 | 75.74 \pm 6.22 | 70.68 \pm 7.79 | 73.61 \pm 10.76 | 75.65 \pm 9.07 |
| Año 4 | 76.56 \pm 7.77 | 77.44 \pm 13.27 | 77.79 \pm 11.91 | 80.38 \pm 10.48 | 82.04 \pm 11.04 |
| Año 5 | 72.89 \pm 7.51 | 77.76 \pm 13.68 | 75.01 \pm 7.91 | 78.94 \pm 11.22 | 74.62 \pm 10.47 |
| Media | 66.04 | 68.47 | 66.63 | 70.26 | 69.15 |

Tabla A.2: Media y desviación típica de los distintos modelos para sensitivity

| Años | Modelos | | | | |
|-------|------------------|------------------|------------------------------------|------------------------------------|------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 1 | 85.60 \pm 4.12 | 91.26 \pm 2.24 | 91.28 \pm 4.49 | 92.16 \pm 1.69 | 89.51 \pm 3.27 |
| Año 2 | 74.10 \pm 5.89 | 80.36 \pm 6.30 | 83.76 \pm 4.90 | 82.85 \pm 6.24 | 78.91 \pm 7.72 |
| Año 3 | 74.15 \pm 5.42 | 80.55 \pm 3.67 | 87.72 \pm 2.19 | 81.40 \pm 3.47 | 79.90 \pm 3.89 |
| Año 4 | 79.16 \pm 5.05 | 83.83 \pm 7.48 | 87.83 \pm 4.50 | 85.01 \pm 5.39 | 85.62 \pm 5.72 |
| Año 5 | 79.51 \pm 2.85 | 85.55 \pm 8.79 | 85.29 \pm 6.05 | 87.19 \pm 10.02 | 86.72 \pm 9.16 |
| Media | 78.51 | 84.31 | 87.09 | 85.72 | 84.13 |

Tabla A.3: Media y desviación típica de los distintos modelos para specificity

| Años | Modelos | | | | |
|--------------|------------------|------------------------------------|------------------------------------|-------------------------------------|------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 1 | 35.59 \pm 5.75 | 42.59 \pm 10.17 | 46.20 \pm 14.01 | 48.89 \pm 10.12 | 40.55 \pm 9.62 |
| Año 2 | 56.48 \pm 8.20 | 63.76 \pm 5.16 | 60.30 \pm 6.55 | 65.93 \pm 4.85 | 64.19 \pm 4.92 |
| Año 3 | 63.99 \pm 6.78 | 71.49 \pm 5.67 | 71.93 \pm 6.51 | 70.54 \pm 6.88 | 71.01 \pm 6.76 |
| Año 4 | 71.71 \pm 4.75 | 76.97 \pm 8.99 | 77.43 \pm 7.10 | 79.32 \pm 7.77 | 80.46 \pm 6.04 |
| Año 5 | 70.31 \pm 4.52 | 78.65 \pm 7.43 | 75.06 \pm 9.38 | 80.65 \pm 6.24 | 77.15 \pm 2.26 |
| Media | 59.62 | 66.69 | 66.18 | 69.07 | 66.67 |

Tabla A.4: Media y desviación típica de los distintos modelos para fl-score

A.2 Resultados utilizando los datos de la visita en el mes 12

En comparación con el experimento anterior, esta prueba destaca por varios aspectos. Primero, se observa una disminución significativa en la diferencia entre los resultados del modelo entrenado con ODE-RGRU_set y los de los demás modelos en prácticamente todas las métricas, aunque sigue siendo el modelo con peores resultados. Por otro lado, el modelo entrenado con ADNIMERGE_set muestra los resultados más robustos y consistentes, contrastando con el primer experimento, donde el modelo entrenado con RSF_set es el que mejores resultados presenta en términos generales. En esta prueba, el modelo entrenado con RSF_set se posiciona como el segundo con mejores resultados.

Por tanto, se puede concluir que el hecho de utilizar los datos de la visita en el mes 12 supone una mejora de los resultados para todas las métricas en los modelos entrenados con ODE-RGRU_set y ADNIMERGE_set respecto a los resultados obtenidos en la prueba anterior. Sin embargo, para los otros tres modelos implica un empeoramiento en los resultados para todas las métricas salvo para sensitivity, donde se percibe una mejora relevante para todos los modelos, y AUC. Asimismo, se observa una alteración significativa para el caso del modelo entrenado con EBM_set cuyos resultados son peores en términos generales al utilizar los datos de la visita en el mes 12. Esta falta de coherencia entre los resultados (el rendimiento de algunos modelos se ve beneficiado y en otros modelos perjudicado) sugiere que este experimento es muy dependiente del conjunto utilizado.

| Años | Modelos | | | | |
|--------------|------------------|------------------|------------------|------------------|------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 2 | 74.69 \pm 3.89 | 79.54 \pm 5.25 | 78.94 \pm 3.45 | 79.89 \pm 3.92 | 81.67 \pm 3.72 |
| Año 3 | 77.22 \pm 6.28 | 81.31 \pm 4.26 | 79.08 \pm 2.51 | 80.05 \pm 5.01 | 83.20 \pm 2.64 |
| Año 4 | 78.77 \pm 3.33 | 78.38 \pm 3.36 | 82.46 \pm 3.86 | 83.32 \pm 3.58 | 83.32 \pm 4.84 |
| Año 5 | 80.83 \pm 5.66 | 83.80 \pm 7.37 | 82.88 \pm 7.68 | 83.27 \pm 5.73 | 83.76 \pm 6.40 |
| Media | 77.85 | 80.75 | 80.84 | 81.63 | 82.99 |

Tabla A.5: Media y desviación típica de los distintos modelos para accuracy

| Años | Modelos | | | | |
|--------------|-------------------------------------|-------------------|-------------------|-------------------|-------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 2 | 62.30 \pm 13.86 | 57.35 \pm 3.28 | 57.00 \pm 1.64 | 54.34 \pm 6.45 | 56.29 \pm 6.57 |
| Año 3 | 78.17 \pm 11.16 | 72.23 \pm 10.22 | 67.96 \pm 7.23 | 70.55 \pm 10.96 | 73.56 \pm 7.44 |
| Año 4 | 75.46 \pm 8.86 | 80.67 \pm 17.25 | 76.22 \pm 12.60 | 82.89 \pm 16.66 | 82.82 \pm 16.66 |
| Año 5 | 79.67 \pm 9.75 | 81.94 \pm 7.27 | 79.07 \pm 17.76 | 80.29 \pm 12.16 | 82.11 \pm 9.41 |
| Media | 73.92 | 73.05 | 70.06 | 72.02 | 73.70 |

Tabla A.6: Media y desviación típica de los distintos modelos para sensitivity

| Años | Modelos | | | | |
|--------------|------------------|-------------------|------------------|-------------------|-------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 2 | 77.21 \pm 5.21 | 84.44 \pm 5.69 | 84.24 \pm 2.71 | 85.53 \pm 4.17 | 87.31 \pm 4.02 |
| Año 3 | 76.53 \pm 6.15 | 84.26 \pm 4.84 | 82.88 \pm 3.80 | 83.74 \pm 3.74 | 86.54 \pm 2.90 |
| Año 4 | 80.27 \pm 5.76 | 78.54 \pm 5.15 | 85.45 \pm 3.14 | 84.80 \pm 5.41 | 84.79 \pm 7.45 |
| Año 5 | 81.58 \pm 5.78 | 84.30 \pm 10.10 | 84.70 \pm 3.81 | 84.37 \pm 11.54 | 84.85 \pm 10.66 |
| Media | 78.90 | 82.88 | 84.32 | 84.61 | 85.87 |

Tabla A.7: Media y desviación típica de los distintos modelos para specificity

| Años | Modelos | | | | |
|--------------|-------------------|-------------------|-------------------|------------------|------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 2 | 46.79 \pm 8.97 | 51.15 \pm 4.72 | 48.69 \pm 8.94 | 49.77 \pm 3.83 | 52.96 \pm 5.01 |
| Año 3 | 63.02 \pm 10.71 | 65.44 \pm 10.69 | 61.60 \pm 5.36 | 63.61 \pm 8.71 | 68.19 \pm 7.49 |
| Año 4 | 67.82 \pm 6.08 | 68.73 \pm 4.59 | 72.03 \pm 5.63 | 74.63 \pm 4.71 | 74.78 \pm 5.82 |
| Año 5 | 73.75 \pm 6.88 | 77.79 \pm 8.23 | 74.90 \pm 12.78 | 76.52 \pm 6.65 | 77.38 \pm 9.01 |
| Media | 62.84 | 65.78 | 64.31 | 66.13 | 68.33 |

Tabla A.8: Media y desviación típica de los distintos modelos para f1-score

A.3 Resultados utilizando dos instantes temporales consecutivos

Al incluir dos instantes temporales consecutivos, se destaca la gran variabilidad que existe entre los distintos años, ya que en ninguna métrica hay un modelo que sea superior en la mayoría de los años. En este caso se destaca accuracy, donde el mejor modelo en media es el entrenado con ODE-RGRU_set, el cual ha sido el modelo con peores resultados en las cinco métricas en las pruebas anteriores realizadas.

Si se comparan los resultados de los modelos cuando se utiliza la visita en baseline y los obtenidos cuando se incluyen dos instantes temporales consecutivos se observan cambios notables. En primer lugar, se aprecia una mejoría en sensitivity en los resultados para todos los modelos. La mejora es especialmente significativa en los modelos entrenados con ODE-RGRU_set y MinimalRNN_set. Para el resto de las métricas, se observa una ligera disminución en los resultados en cuatro de los modelos analizados. El único modelo que experimenta una mejoría en todas las métricas es el entrenado con ODE-RGRU_set. Por último, la métrica que muestra más variación dependiendo del modelo es AUC. Para los modelos entrenados con ODE-RGRU_set y EBM_set, se observa una mejora, mientras que para el modelo entrenado con RSF_set los resultados son muy similares, y en el caso de los modelos entrenados con MinimalRNN_set y ADNIMERGE_set, los resultados empeoran ligeramente al utilizar dos instantes temporales consecutivos.

Finalmente, se observa cierta tendencia general al empeoramiento de los resultados en la mayor parte de las métricas al incluir dos instantes temporales consecutivos salvo para el modelo entrenado con ODE-RGRU_set. Solo se da una mejora en sensitivity para todos los modelos. Este escenario es similar para los modelos entrenados con EBM_set y RSF_set al observado cuando se utilizan los datos de la visita del mes 12, ya que supone un empeoramiento generalizado del rendimiento del modelo.

Cuando se comparan los resultados de los modelos que incluyen dos instantes temporales consecutivos frente a los obtenidos cuando se utiliza la visita en el mes 12, en términos generales no se perciben cambios significativos en el rendimiento. Sin embargo, se pueden destacar pequeñas diferencias. Para el modelo entrenado con ODE-RGRU_set, la inclusión de dos instantes temporales consecutivos no implica una mejora en los resultados obtenidos, salvo en specificity donde es ligeramente superior, sino que se observa una tendencia general al empeoramiento con resultados inferiores, en comparación con el modelo entrenado con los datos de la segunda visita, en cuatro de las cinco métricas.

En el caso del modelo entrenado con `MinimalRNN_set` se observa una mejora de los resultados en prácticamente todas las métricas, destacando principalmente los resultados en specificity siendo el mejor resultado de todos los obtenidos en las diversas pruebas realizadas, lo mismo ocurre en la métrica f1-score pero en este caso la diferencia es bastante más pequeña. El modelo entrenado con `EBM_set` muestra una muy ligera mejora en todas las métricas cuando se utilizan dos instantes consecutivos destacando la mejora en AUC. La métrica donde no se observa mejora es sensitivity, con una pequeña disminución al utilizar dos instantes consecutivos.

Para el modelo entrenado con `RSF_set`, se experimentan mejoras en sensitivity y f1-score. Sin embargo, los resultados para accuracy y specificity son peores al usar dos instantes consecutivos. A pesar de esta disminución, el grado de mejora en otras métricas es mayor que el grado de empeoramiento en estas. Para el AUC, los resultados permanecen constantes en media. Finalmente, el modelo entrenado con `ADNIMERGE_set` muestra un empeoramiento notable de los resultados en todas las métricas cuando se utilizan dos instantes consecutivos, lo cual es significativo, ya que en el experimento del apartado 3.2 sí que se observaba una tendencia de mejora general en el rendimiento cuando se utilizan los datos de la visita en el mes 12. Este resultado sugiere que, para este conjunto, la inclusión de datos de la visita de baseline no beneficia el rendimiento del modelo y, de hecho, puede ser perjudicial.

En conclusión, no se han observado mejoras significativas al utilizar dos instantes temporales consecutivos. Además, al igual que ocurría en el apartado anterior, se detecta una falta de coherencia entre los resultados para los diferentes modelos: para algunos modelos, la inclusión de dos instantes consecutivos mejora relativamente todas las métricas, mientras que para otros, los resultados empeoran o dependen de la métrica evaluada. Esta variabilidad indica que la utilidad de agregar datos de múltiples instantes puede ser altamente dependiente del conjunto de datos específico y las métricas consideradas.

| Años | Modelos | | | | |
|-------|---------------------------|-----------------------------|----------------------|----------------------|----------------------------|
| | <code>ODE-RGRU_set</code> | <code>MinimalRNN_set</code> | <code>EBM_set</code> | <code>RSF_set</code> | <code>ADNIMERGE_set</code> |
| Año 2 | 79.48 ± 4.38 | 78.48 ± 3.77 | 80.43 ± 1.20 | 79.88 ± 2.88 | 78.07 ± 4.58 |
| Año 3 | 78.86 ± 5.28 | 80.28 ± 4.50 | 80.43 ± 1.20 | 79.88 ± 2.88 | 78.07 ± 4.58 |
| Año 4 | 78.86 ± 5.28 | 80.28 ± 4.54 | 80.35 ± 2.87 | 78.62 ± 7.10 | 78.39 ± 3.06 |
| Año 5 | 81.69 ± 4.71 | 83.49 ± 6.70 | 81.31 ± 10.00 | 83.48 ± 4.82 | 82.87 ± 7.26 |
| Media | 81.41 | 81.25 | 81.17 | 81.09 | 80.79 |

Tabla A.9: Media y desviación típica de los distintos modelos para accuracy

| Años | Modelos | | | | |
|-------|---------------------------|-----------------------------|----------------------|----------------------|----------------------------|
| | <code>ODE-RGRU_set</code> | <code>MinimalRNN_set</code> | <code>EBM_set</code> | <code>RSF_set</code> | <code>ADNIMERGE_set</code> |
| Año 2 | 59.25 ± 10.31 | 67.55 ± 8.21 | 54.00 ± 11.86 | 65.49 ± 5.95 | 54.30 ± 13.64 |
| Año 3 | 70.09 ± 9.54 | 75.52 ± 5.93 | 70.72 ± 7.91 | 69.62 ± 11.26 | 71.06 ± 13.48 |
| Año 4 | 83.45 ± 5.08 | 78.12 ± 5.02 | 75.52 ± 7.66 | 80.70 ± 4.51 | 81.36 ± 4.32 |
| Año 5 | 79.57 ± 6.58 | 84.25 ± 11.85 | 79.86 ± 19.66 | 80.89 ± 12.16 | 77.69 ± 13.94 |
| Media | 73.09 | 76.36 | 70.02 | 74.18 | 71.10 |

Tabla A.10: Media y desviación típica de los distintos modelos para sensitivity

| Años | Modelos | | | | |
|-------|---------------------------|-----------------------------|----------------------|----------------------|----------------------------|
| | <code>ODE-RGRU_set</code> | <code>MinimalRNN_set</code> | <code>EBM_set</code> | <code>RSF_set</code> | <code>ADNIMERGE_set</code> |
| Año 2 | 84.75 ± 5.31 | 81.26 ± 4.08 | 86.41 ± 3.08 | 83.78 ± 4.84 | 84.21 ± 5.12 |
| Año 3 | 82.25 ± 4.50 | 82.26 ± 5.11 | 83.61 ± 2.75 | 82.23 ± 5.65 | 81.29 ± 1.91 |
| Año 4 | 86.87 ± 8.55 | 85.49 ± 8.42 | 85.50 ± 6.53 | 82.88 ± 8.74 | 85.47 ± 6.60 |
| Año 5 | 82.51 ± 5.45 | 83.69 ± 4.34 | 84.02 ± 12.73 | 85.59 ± 4.89 | 86.07 ± 5.94 |
| Media | 84.09 | 83.17 | 84.89 | 83.62 | 84.26 |

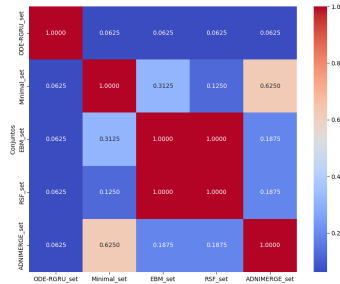
Tabla A.11: Media y desviación típica de los distintos modelos para specificity

| Años | Modelos | | | | |
|--------------|------------------------------------|-------------------------------------|-------------------|-------------------------------------|-------------------------------------|
| | ODE-RGRU_set | MinimalRNN_set | EBM_set | RSF_set | ADNIMERGE_set |
| Año 2 | 53.60 \pm 9.62 | 55.52 \pm 10.15 | 49.65 \pm 6.80 | 56.50 \pm 6.20 | 49.57 \pm 10.00 |
| Año 3 | 64.64 \pm 8.15 | 68.04 \pm 5.97 | 64.33 \pm 5.34 | 64.38 \pm 10.42 | 63.96 \pm 7.96 |
| Año 4 | 80.21 \pm 5.71 | 75.56 \pm 6.33 | 71.99 \pm 10.95 | 75.89 \pm 4.25 | 77.35 \pm 5.71 |
| Año 5 | 76.32 \pm 7.81 | 79.26 \pm 8.93 | 73.63 \pm 13.06 | 78.38 \pm 7.31 | 76.25 \pm 13.03 |
| Media | 68.69 | 69.59 | 64.90 | 68.79 | 66.78 |

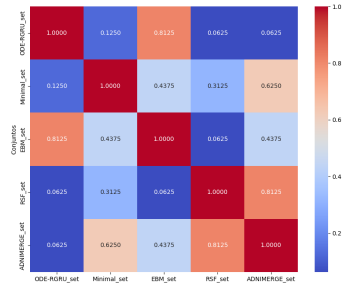
Tabla A.12: Media y desviación típica de los distintos modelos para f1-score

Anexo B: Test de Wilcoxon

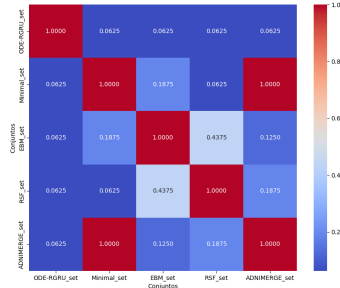
Se ha realizado el test de Wilcoxon para comprobar, utilizando una base estadística, si existen diferencias significativas entre los resultados de la visita inicial utilizando distintos conjuntos de features. En este test, si el valor p es menor que 0.05, se rechaza la hipótesis nula de que no hay diferencia entre las medias y se puede concluir que sí existe una diferencia significativa. Si el valor p es mayor que 0.05, no se puede concluir que existe una diferencia significativa. En este caso, el p-valor más pequeño es 0.06, para ODE-RGRU_set, por lo que no se pueden concluir diferencias significativas en ningún caso. No obstante, este valor p indica que la diferencia observada está cerca del umbral de significación estadística, lo que sugiere la posibilidad de una diferencia que podría ser relevante, aunque no concluyente bajo el criterio estándar de 0.05. La Figura B.1 muestra los valores p de las pruebas de suma de rangos de Wilcoxon de cola derecha por pares para la distribución de los valores obtenidos en el apartado 3.1 para cada una de las métricas.



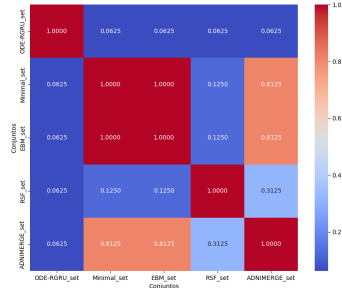
(a) Test de Wilcoxon para accuracy



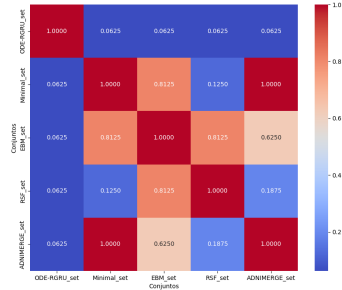
(b) Test de Wilcoxon para sensitivity



(c) Test de Wilcoxon para specificity



(d) Test de Wilcoxon para f1-score



(e) Test de Wilcoxon para AUC

Figura B.1: Test de Wilcoxon para las diferentes métricas

Anexo C: Explicabilidad global complementaria

C.1 Explicabilidad global para la visita en baseline

Como se puede observar en la Figura C.1, en el conjunto de features ODE-RGRU, la feature más relevante es ADAS13, mientras que MMSE es menos relevante en los años 3 y 5. Entre las features anatómicas, la más destacada es MidTemp, con una relevancia similar a lo largo de todos los años. Fusiform también es importante, especialmente a partir del tercer año. Destaca principalmente la mayor influencia de las features de medidas de evaluación neuropsicológica respecto a las features anatómicas.

En cuanto a la explicabilidad global de MinimalRNN_set, como se ilustra en la Figura C.2, se destacan FAQ y FDG como las features más importantes. RAVLT_immediate es relevante en los tres primeros años, pero nada relevante en los dos últimos años. CDRSB es muy relevante en el primer y quinto año, pero tiene poca relevancia en los años 2 y 4. Asimismo MMSE y ADAS13 también tienen bastante relevancia. Entre las features anatómicas destaca Entorhinal con una relevancia media en prácticamente todos los años.

Según lo mostrado en la Figura C.3, la feature con mayor influencia en la predicción a lo largo de los años para EBM_set es FAQ. También se destaca la relevancia de APOE ϵ 4, a pesar de que en el último año la influencia sea menor que en el resto. Por otro lado, sorprende la alta influencia del género (PTGENDER) en las predicciones, relevante en todos los años salvo en el cuarto. Asimismo, subrayar la relevancia de ADAS13, CDRSB y MMSE. El primero de ellos muy relevante principalmente en los 4 últimos años, CDRSB destaca por su importancia en la predicción en el primer y el último año, mientras que en el cuarto su importancia disminuye, y MMSE muy relevante en los 4 primeros años. En cuanto a las features de medidas MRI, cabe destacar que la influencia de las mismas varía dependiendo del año, por ejemplo, Entorhinal presenta influencia notable en los años 1, 3 y 4; sin embargo, en los años 2 y 5 su relevancia es bastante inferior. Destaca MidTemp con una relevancia significativa en el cuarto año, y con menor importancia en términos generales se encontrarían Entorhinal y Fusiform.

Como se puede percibir en la Figura C.4, en ADNIMERGE_set se observa cierta variabilidad en la relevancia de las features a lo largo de los distintos años. Entre las features más importantes, destaca nuevamente FAQ como la más relevante. FDG también es muy relevante en todos los años excepto en el primero. CDRSB cobra bastante importancia en los años 1, 3 y 5, mientras que su relevancia disminuye considerablemente en los otros dos años. APOE ϵ 4 es bastante relevante en todos los años, excepto en el último. Por otro lado, PTGENDER es menos relevante en comparación con EBM_set, teniendo relativa importancia en los años 1, 4 y 5. Finalmente, entre las features anatómicas, de nuevo la más relevante es Entorhinal aunque la relevancia de este grupo de features no es especialmente destacable. De igual forma que ocurría en MinimalRNN_set, RAVLT_immediate es relevante en los tres primeros años aunque en los dos últimos apenas influye en la predicción.

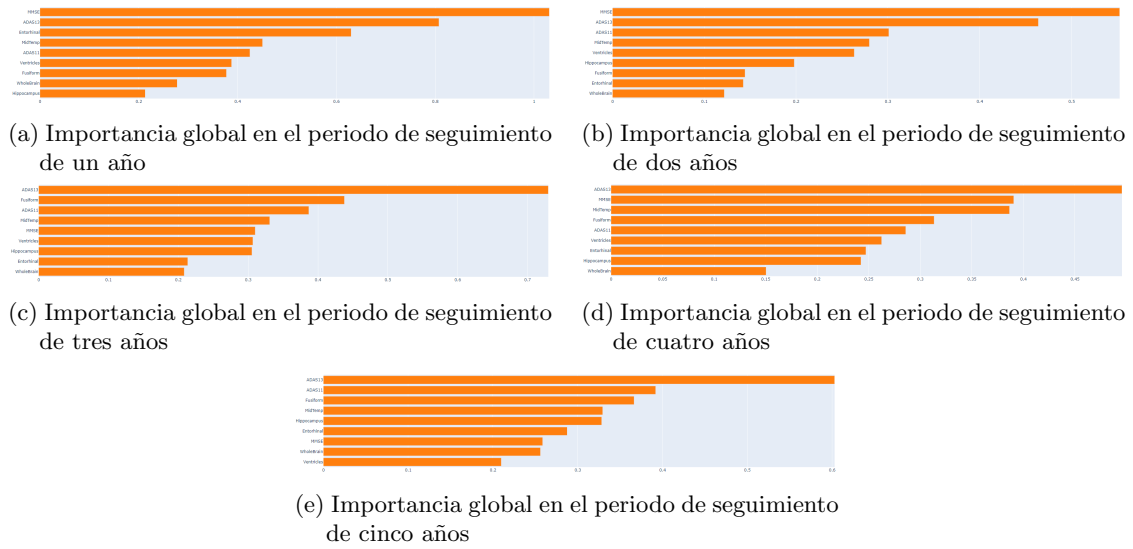


Figura C.1: Modelo ODE-RGRU_set. Explicabilidad global en baseline

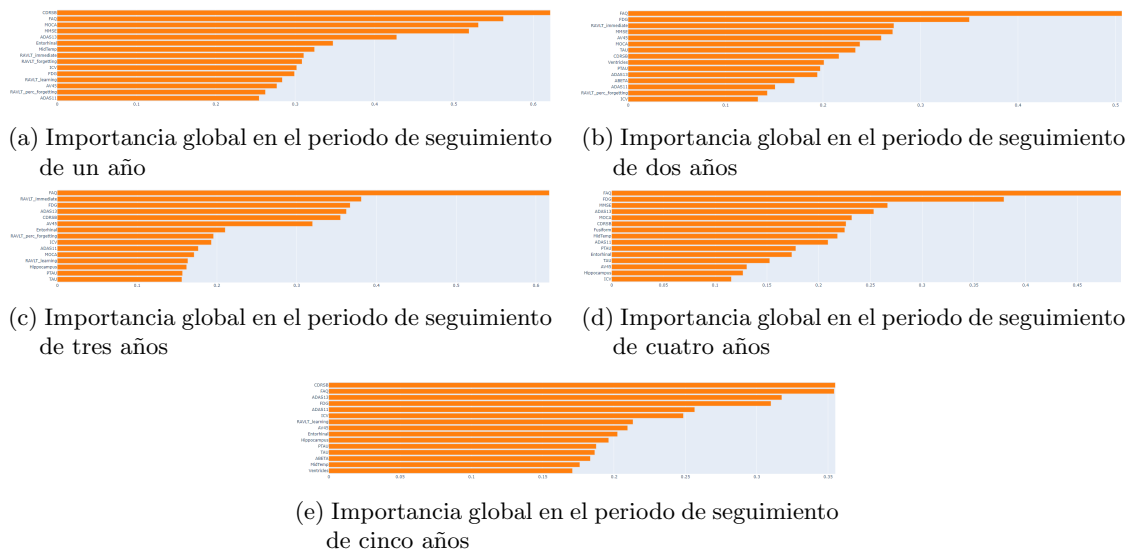


Figura C.2: Modelo MinimalRNN_set. Explicabilidad global en baseline

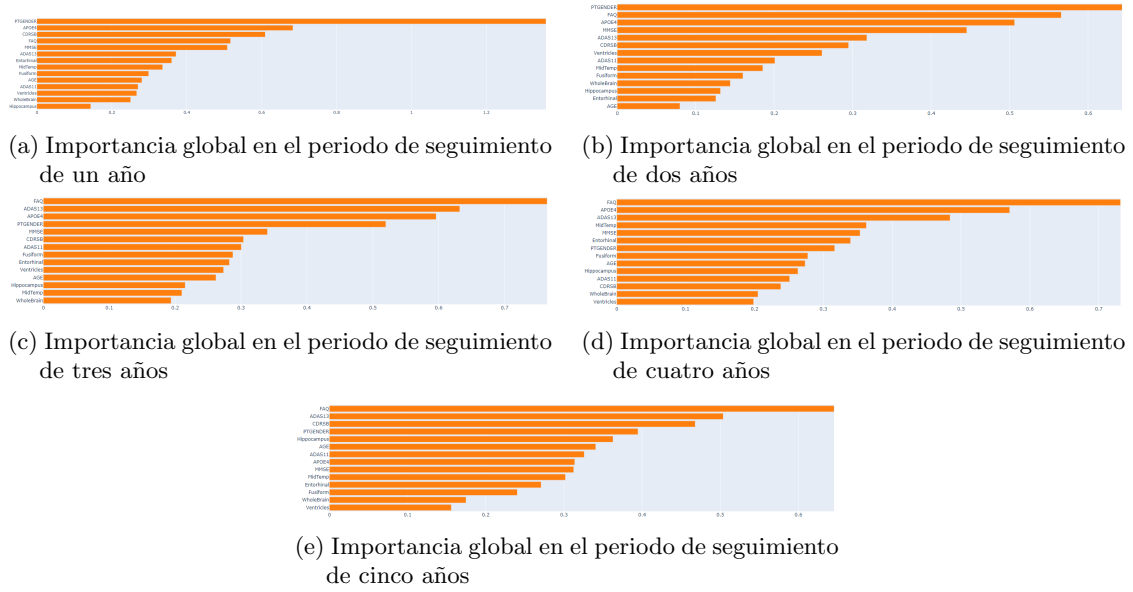


Figura C.3: Modelo EBM_set. Explicabilidad global en baseline

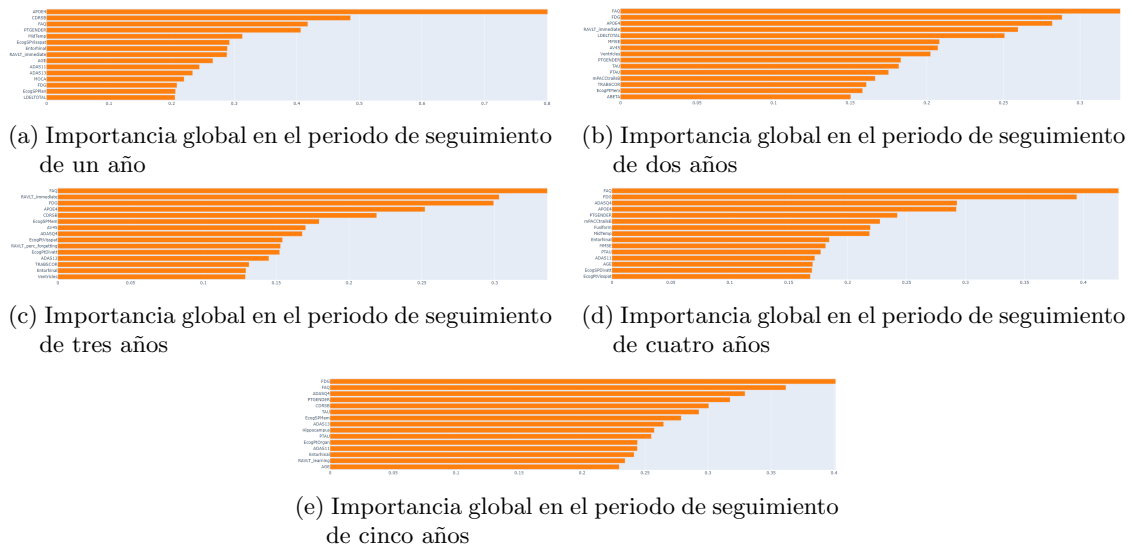


Figura C.4: Modelo ADNIMERGE_set. Explicabilidad global en baseline

C.2 Explicabilidad global para la visita en el mes 12

Como se puede advertir en la Figura C.5, en ODE-RGRU_set al utilizar los datos de la segunda visita para realizar la predicción, se observa que las features de evaluación neuropsicológica (ADAS11, ADAS13 y MMSE) son más relevantes en términos generales que las features anatómicas, ya que son las tres más importantes en todos los años salvo en el año 4. Por otro lado, MidTemp pierde relevancia, especialmente en los dos primeros años. La feature anatómica más relevante en términos generales es Hippocampus.

En la Figura C.6, se observa que en el caso de MinimalRNN_set, FAQ sigue siendo la feature más relevante, a pesar de que en el cuarto año su relevancia disminuye. Destaca la escasa relevancia de FDG, importante únicamente en el último año. En contrapunto, MMSE y ADAS13 aumentan su relevancia en comparación con la visita inicial. En el año 4 se produce una variabilidad notable en la importancia de las features, donde son las features anatómicas las más relevantes. Finalmente, RAVLT_immediate mantiene su relevancia y CDRSB presenta mayor importancia en comparación con la visita inicial.

En la Figura C.7, se muestra la explicabilidad global del conjunto EBM_set. En términos generales, la relevancia de las features es similar a la observada en la primera visita, aunque existen algunos cambios. FAQ es menos importante, destacando su baja relevancia en el año 4. APOE ϵ 4 es la feature más importante en el año 4, pero poco relevante en los años 3 y 5. PTGENDER sigue siendo relevante, pero su importancia disminuye en los dos últimos años. MMSE mantiene una relevancia similar a la observada inicialmente. CDRSB es muy poco relevante en el año 2, pero es muy relevante en los dos últimos años. De nuevo, el año 4 se caracteriza por su gran variabilidad y una mayor importancia de las features anatómicas.

En el caso de RSF_set, como se observa en la Figura C.8, las features más relevantes en la predicción son FAQ y FDG. Sin embargo, destaca que ambas features son muy poco o nada relevantes en el tercer año. Por otro lado, también destacan las features APOE ϵ 4 y CDRSB, importantes en la visita de baseline pero que en este caso, su importancia depende del año. En el caso de APOE ϵ 4, en los años 2 y 4 es muy relevante, mientras que en los años 3 y 5 es menos relevante. En cuanto a CDRSB, es muy relevante en los dos últimos años y poco relevante en los dos primeros. Se destaca la importante relevancia de las features anatómicas en el caso del año 4, donde MidTemp, Hippocampus, Entorhinal y WholeBrain tienen una importancia superior a FAQ, el resto de los años, las features de este tipo son muy poco relevantes, únicamente Hippocampus en el año 5. Por otro lado, RAVLT_immediate es muy relevante en los años 2 y 3 y nada relevante para los dos últimos años. Asimismo, PTGENDER tiene una relevancia inferior en comparación a la explicabilidad en baseline en este conjunto, destacando su nula relevancia en el año 5.



Figura C.5: Modelo ODE-RGRU_set. Explicabilidad global en el mes 12



C.3 Explicabilidad global utilizando dos instantes temporales consecutivos

En primer lugar para ODE-RGRU_set, como se observa en la Figura C.9, destaca de nuevo la importancia de los test cognitivos frente a las features anatómicas. No obstante, en el año 4 se observa un incremento en la relevancia de este tipo de features. Además, en este conjunto, se aprecia la importancia de las features en el mes 12, lo cual también es propiciado por ser un conjunto con un número reducido de features.

Para el conjunto MinimalRNN_set, las features más importantes son nuevamente FDG y FAQ, como se muestra en la Figura C.10. En el caso de FDG, el valor en baseline es más importante que en la segunda visita para todos los años, salvo el cuarto. En el caso de FAQ, la importancia varía dependiendo del año. De nuevo, RAVLT_immediate es importante, siendo relevante en m12 para los dos primeros años, mientras que en baseline es relevante únicamente en el tercer año. Finalmente, CDRSB destaca por su relevancia en los dos últimos años.

En RSF_set, FAQ es la feature más importante tanto en baseline como en m12, y su relevancia varía dependiendo del año: en el año 2, ambos son muy importantes; en los años 3 y 4, FAQ_m12 es más relevante; y en el último año, FAQ en baseline vuelve a ser más importante. Como se observa en la Figura C.11, destaca la escasa importancia de FDG, que solía ser muy relevante en la explicabilidad de los conjuntos de features que la incluían. De nuevo, RAVLT_immediate es importante, siguiendo el mismo patrón que en MinimalRNN_set. APOE ϵ 4, tanto en baseline como en m12, es importante en los años 2 y 4. Además, ADAS11 y ADAS13 en baseline son relevantes en los años 3 y 5.

En el caso de ADNIMERGE_set, como se muestra en la Figura C.12, ADASQ4 destaca por su importancia en los dos primeros años. Asimismo, RAVLT_immediate presenta el mismo comportamiento que se da en MinimalRNN_set y RSF_set. También se observa la escasa importancia de FDG, como en el conjunto anterior, y de FAQ, cuya relevancia disminuye, principalmente en el año 5. Finalmente, mPACCdigit es relevante en los dos primeros años, de la misma forma que ocurría en la explicabilidad de este conjunto en el mes 12.

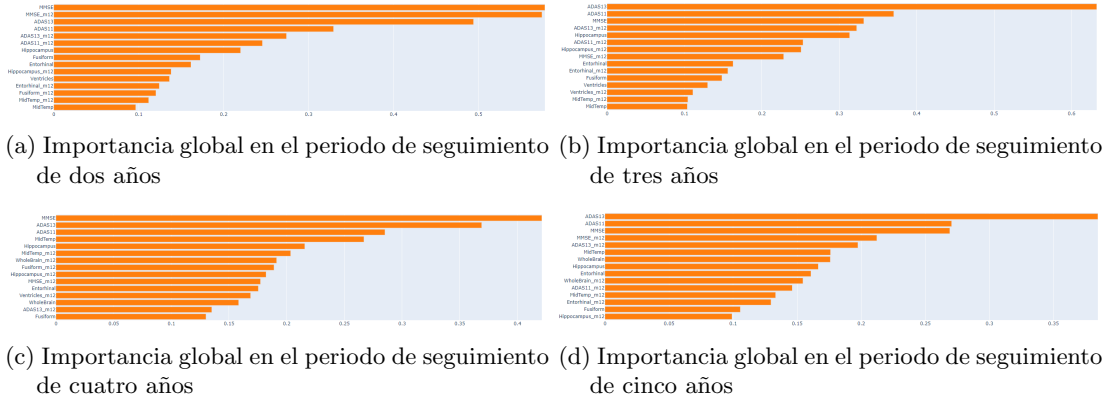


Figura C.9: Modelo ODE-RGRU_set. Explicabilidad global utilizando dos instantes consecutivos



Figura C.10: Modelo MinimalRNN_set. Explicabilidad global utilizando dos instantes consecutivos

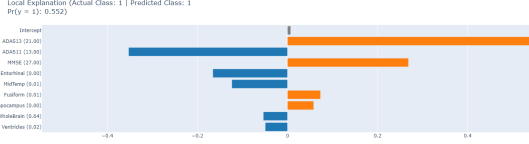


Figura C.11: Modelo RSF_set. Explicabilidad global utilizando dos instantes consecutivos

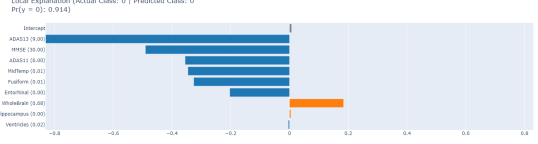


Figura C.12: Modelo ADNIMERGE_set. Explicabilidad global utilizando dos instantes consecutivos

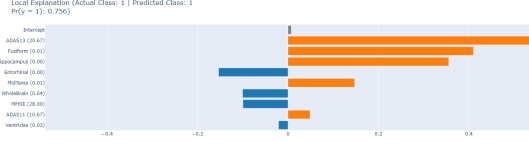
Anexo D: Explicabilidad local complementaria



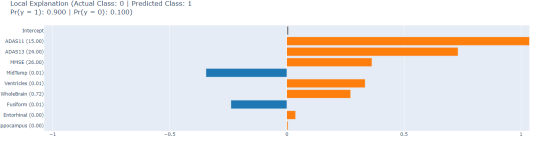
(a) Importancia de las features para el paciente 1



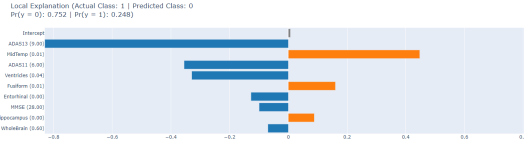
(b) Importancia de las features para el paciente 2



(c) Importancia de las features para el paciente 3

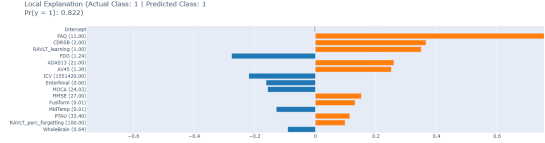


(d) Importancia de las features para el paciente 4

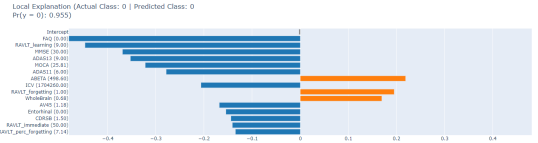


(e) Importancia de las features para el paciente 5

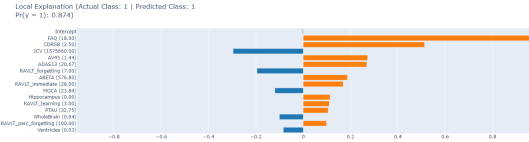
Figura D.1: Modelo ODE-RGRU_set. Explicabilidad local en baseline de 5 pacientes aleatorios



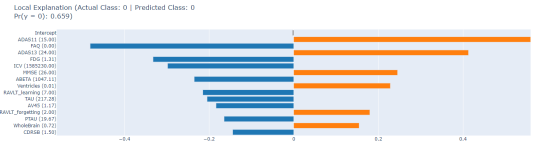
(a) Importancia de las features para el paciente 1



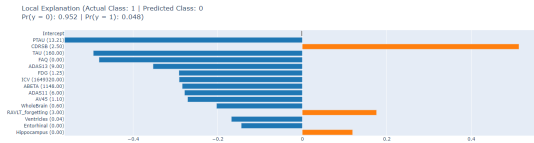
(b) Importancia de las features para el paciente 2



(c) Importancia de las features para el paciente 3



(d) Importancia de las features para el paciente 4



(e) Importancia de las features para el paciente 5

Figura D.2: Modelo MinimalRNN_set. Explicabilidad local en baseline de 5 pacientes aleatorios

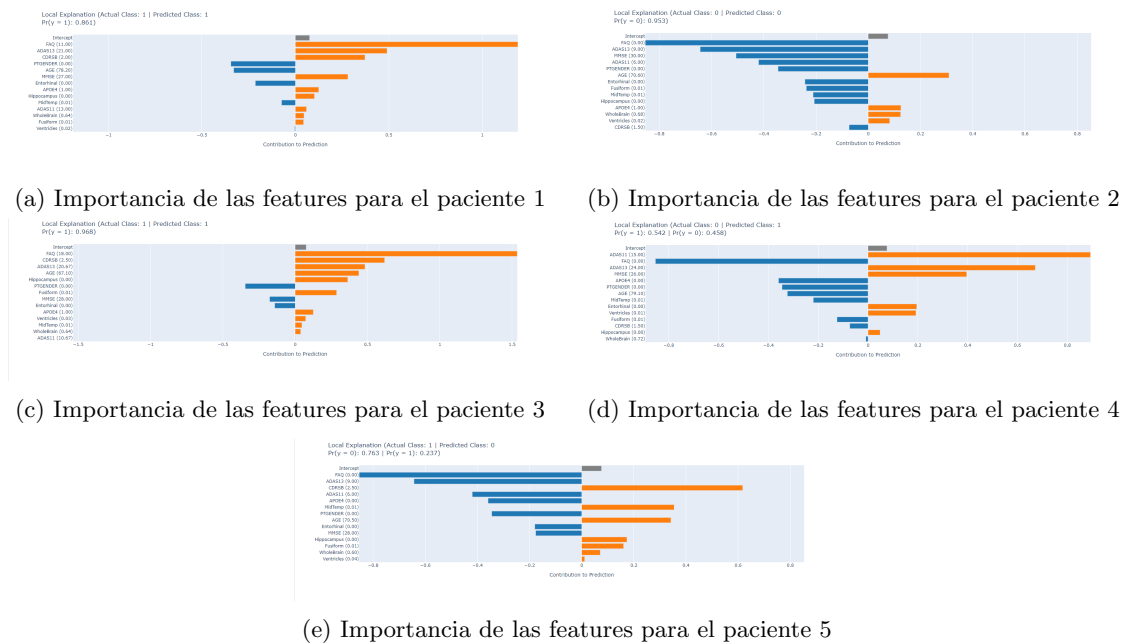


Figura D.3: Modelo EBM_set. Explicabilidad local en baseline de 5 pacientes aleatorios

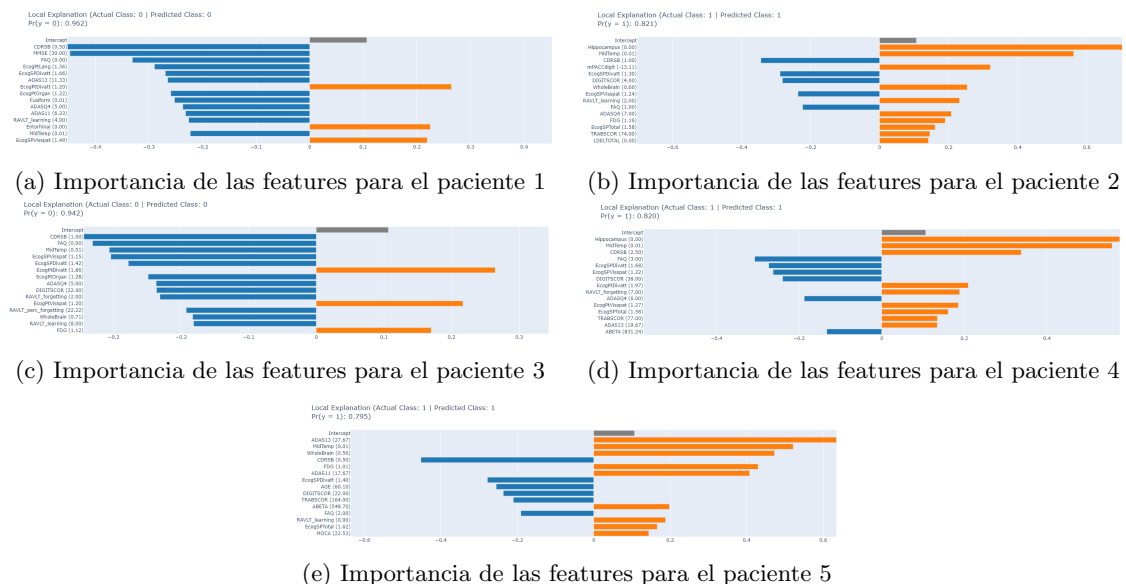


Figura D.4: Modelo ADNIMERGE_set. Explicabilidad local en baseline de 5 pacientes aleatorios

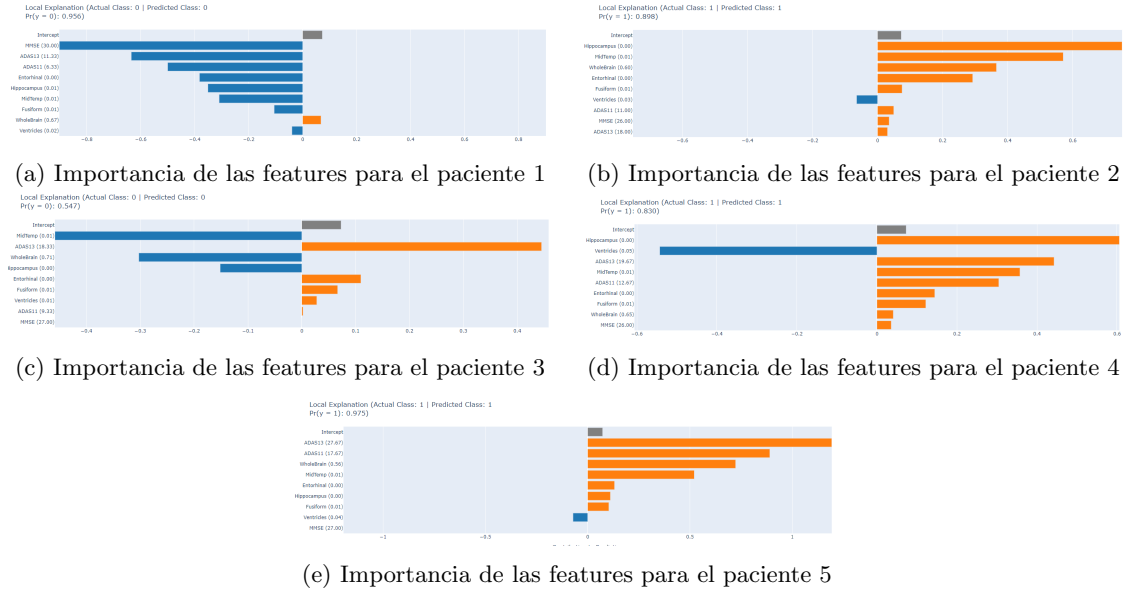


Figura D.5: Modelo ODE-RGRU_set. Explicabilidad local en el mes 12 de 5 pacientes aleatorios

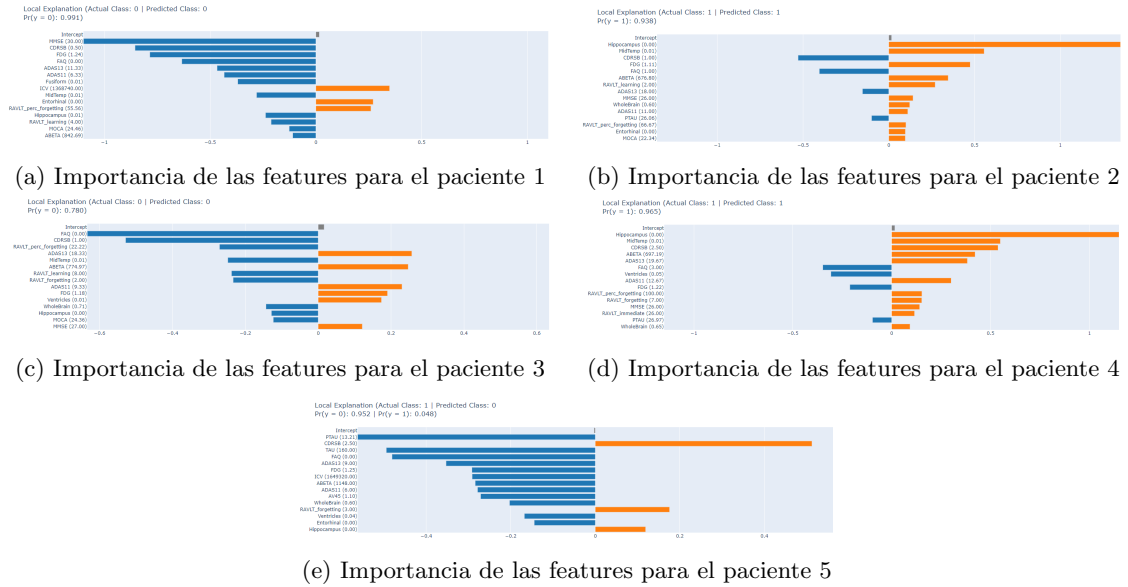


Figura D.6: Modelo MinimalRNN_set. Explicabilidad local en el mes 12 de 5 pacientes aleatorios

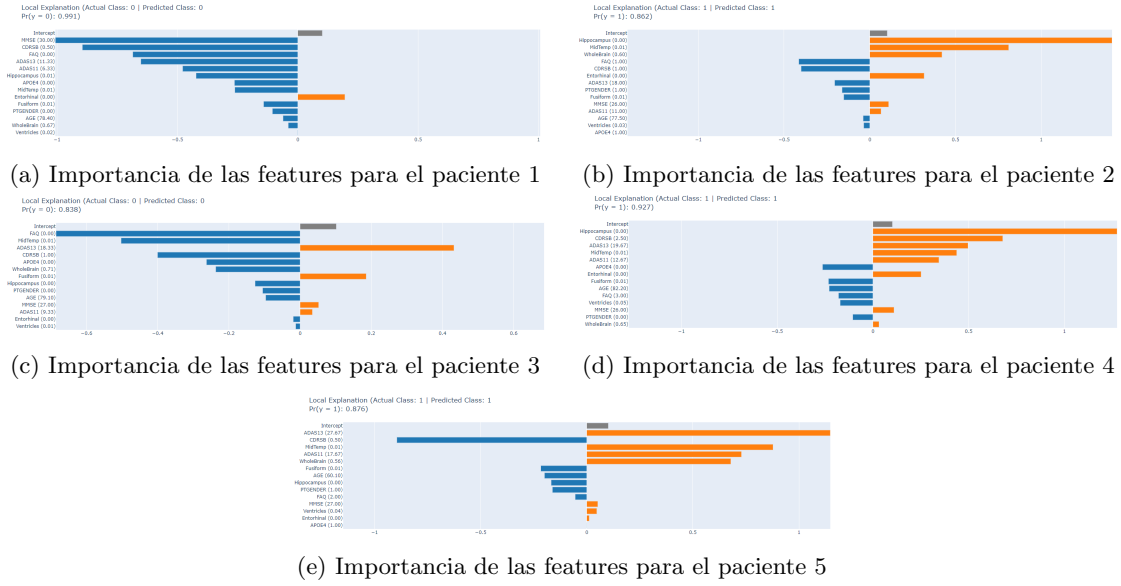


Figura D.7: Modelo EBM_set. Explicabilidad local en baseline de 5 pacientes aleatorios

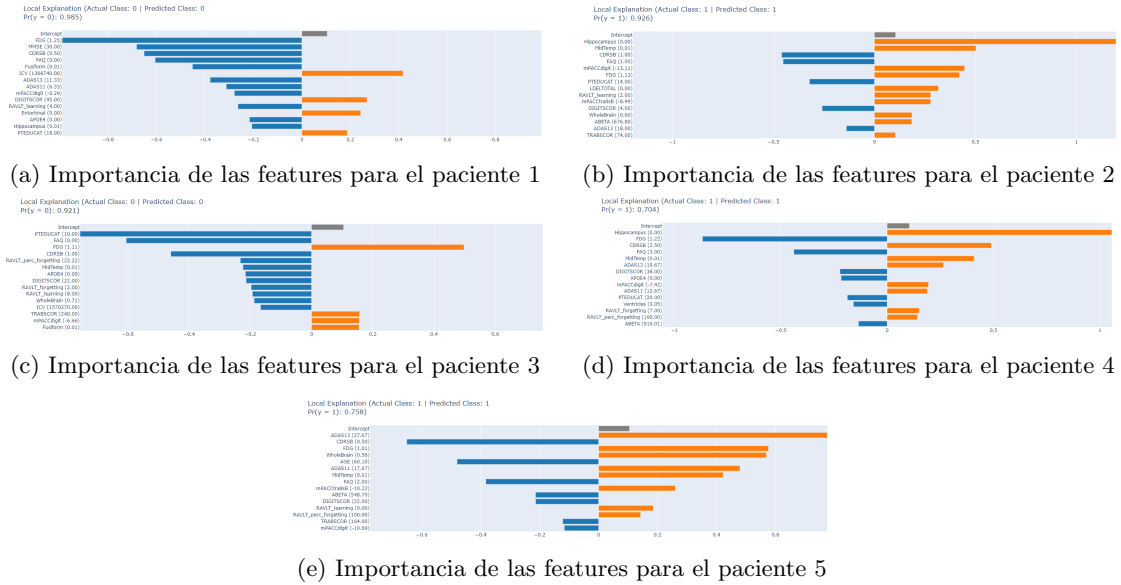


Figura D.8: Modelo RSF_set. Explicabilidad local en el mes 12 de 5 pacientes aleatorios

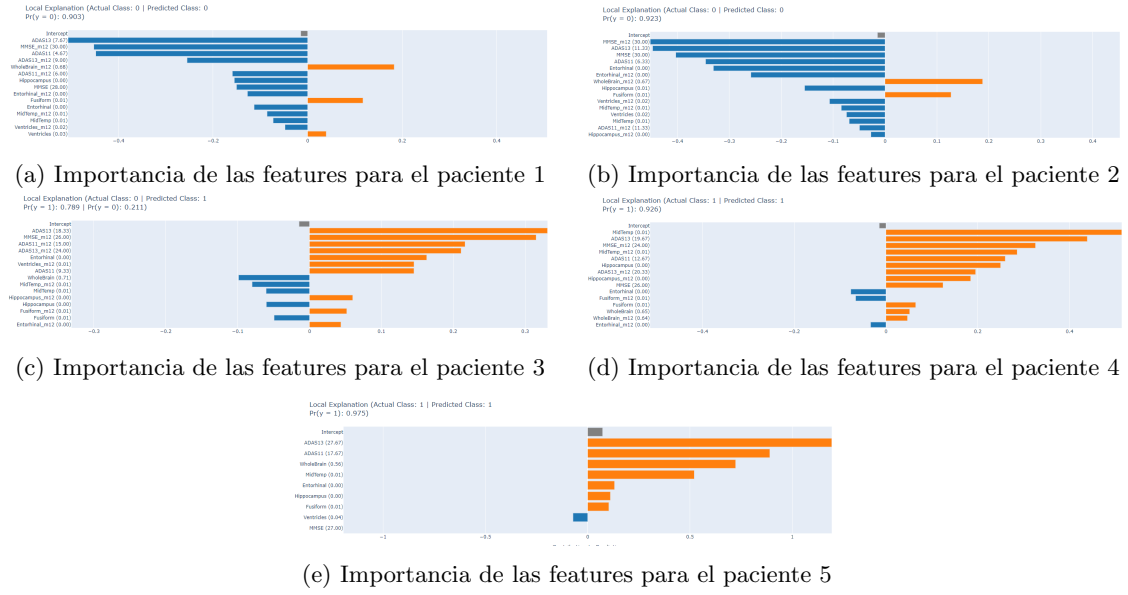


Figura D.9: Modelo ODE-RGRU_set. Explicabilidad local utilizando dos instantes temporales consecutivos de 5 pacientes aleatorios

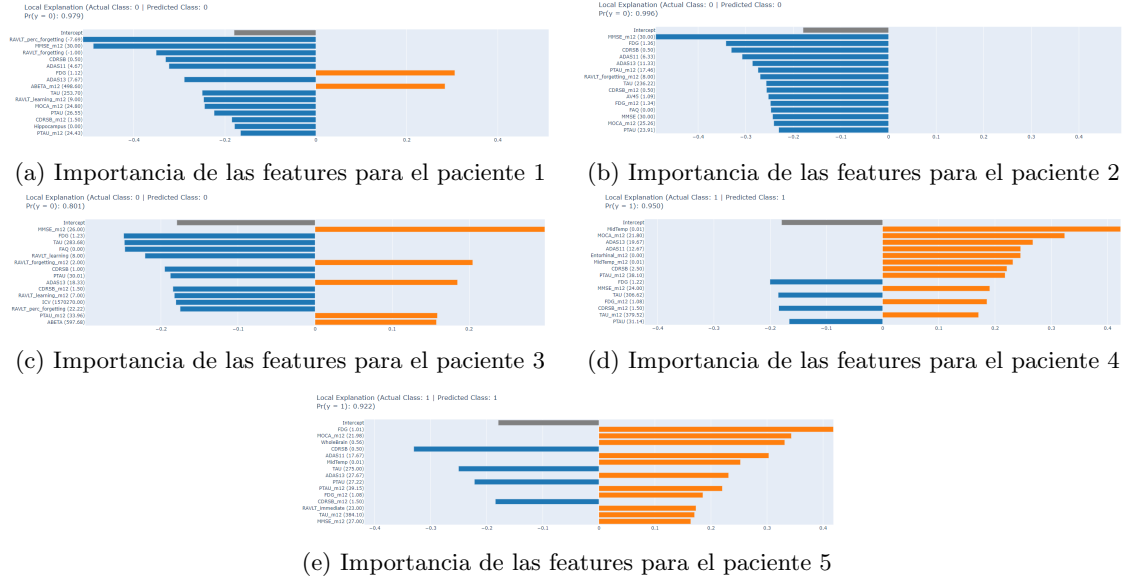


Figura D.10: Modelo MinimalRNN_set. Explicabilidad local utilizando dos instantes temporales consecutivos de 5 pacientes aleatorios

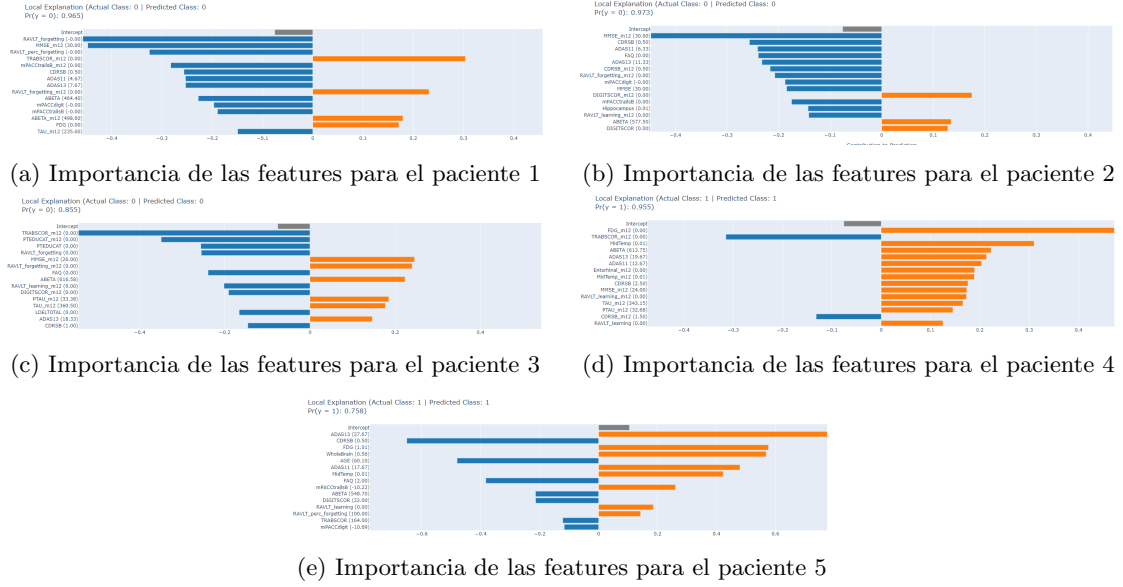


Figura D.11: Modelo RSF_set. Explicabilidad local utilizando dos instantes temporales consecutivos de 5 pacientes aleatorios

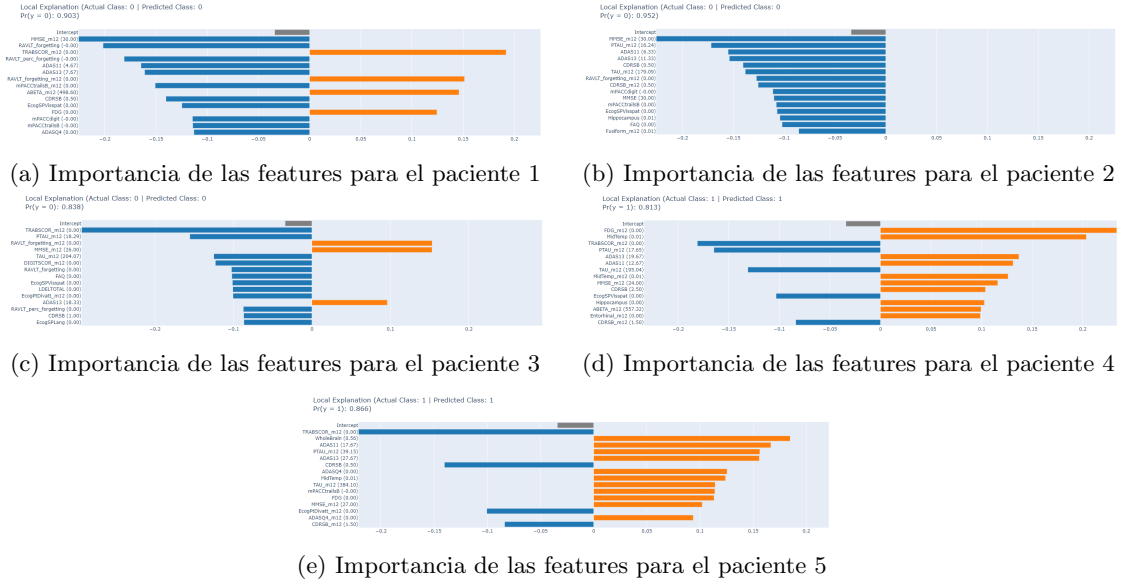


Figura D.12: Modelo ADNIMERGE_set. Explicabilidad local utilizando dos instantes temporales consecutivos de 5 pacientes aleatorios